THE EMIGRANT LETTER DIGITISED: MARKUP AND ANALYSIS

by

EMMA LOUISE MORETON

A thesis submitted to the University of Birmingham for the degree of DOCTOR OF PHILOSOPHY

Department of English Language and Applied Linguistics
University of Birmingham
September 2015

ABSTRACT

The sourcing, preservation and documentation of emigrant letter collections is now gathering pace, with the Internet providing a significant new forum for the dissemination of long-hidden archives. Most existing digital letter collections consist of unannotated versions of original manuscripts. The digitisation process has made the letters more accessible and has also increased their searchability. However, relatively few emigrant letter projects have moved beyond the digitisation stage to exploit text content and enhance usability and searchability through the use of digital technologies.

This thesis explores some of the opportunities and challenges of working with digitised historical emigrant letter collections. Essentially, the thesis does two things: first, it uses digital technologies (specifically corpus and computational methods of analysis) to explore the language of emigrant letters, building on the existing body of research – primarily by historians – to offer another way into migrant correspondence; second, it proposes a system of markup for capturing metadata relating to emigrant letters – metadata which can then be used to interconnect resources enabling users to carry out more nuanced and sophisticated searchers. I argue that my proposed system could be widely applied to emigrant letter collections, facilitating much greater interdisciplinary and collaborative analysis of such material than has been undertaken hitherto.

ACKNOWLEDGEMENTS

There are two people in particular that I would like to acknowledge. The first person is Patrick O'Sullivan whom I contacted many years ago with a vague idea of using corpus methods to examine nineteenth century Irish emigrant letters. Paddy put me in contact with the right people, regularly forwarded useful papers and details of projects, and generally kept me informed on things that were relevant to my research. The second person that I would like to acknowledge is Kerby Miller. In 2009, and again in 2010, with funding from Coventry University, I had the pleasure of visiting Professor Miller in Missouri to work with his archive of Irish emigrant correspondence. Kerby gave me unlimited access to his collection of letters, suggested interesting letter series I might want to explore, helped with the practicalities of copying and transcribing the letters, and provided huge amounts of contextual information relating to the different authors. He is a generous, knowledgeable and inspiring scholar. Enough said!

Over the course of the past few years I have presented my work at various conferences and published essays in various journals. I owe a lot to the conference participants, journal editors and peer reviewers who gave me their time, suggestions and feedback. I would like to give a special mention to Tony Fairman, who always had encouraging words to say. It is not possible to list names – obviously – but the publication / peer review process, in particular, has helped me enormously in developing my ideas, giving me the confidence to explore other disciplines and broaden my field of research. I hope that I can pass this very positive experience onto others.

I would also like to acknowledge both sets of parents: Vicky and Peter Moreton and Wendy and John Cummin. Without their very practical support I would not have been able to finish this. And, finally, I want to thank my partner – Dan Cummin – for always being there and for helping to keep me on track.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ALC	Alice Lough Corpus
ELC	Elizabeth Lough Corpus
JLC	Julia Lough Corpus
NLC	Annie Lough Corpus

INTRODUCTION

News from America by James Brennan (1837-1907)[1]

Over the past decade in particular, there has been a marked increase in the creation and development of historical letter corpora. This has been prompted in part by a renewed recognition that such material can reflect aspects of the spoken language of those distant times. ‘Correspondence often resembles spoken registers more closely than other types of writing’, Nevalainen and Raumolin-Brunberg note (1996, p. 39) citing Biber (1995, pp. 283-300), thereby providing linguists with a window into how language was used at a particular period in time, as well as providing insight into the letter writers themselves – their experiences, preoccupations and beliefs – and the historical context in which they wrote. However, as Auer and Fairman point out, many of these historical letter corpora contain samples of writing by the classically educated members of society (2012, pp. 77-78). For example, the *Mapping the Republic of Letters*

[1] Brennan, J. (1837-1907) News from America. Available from http://www.crawfordartgallery.ie/pages/paintings/JamesBrennan.html [Accessed 30 July 2015].

project at Stanford University, in collaboration with various partners including Oxford University's *Cultures of Knowledge* project,[2] maps networks of correspondence between 'scientific academies', within Europe and America during, primarily, the seventeenth and eighteenth centuries, to explore how such networks facilitated, amongst other things, the development and dissemination of ideas and the spread of political news.[3] The *Darwin Correspondence Project* at Cambridge University has collected and digitised roughly fifteen thousand letters by Charles Darwin (to be published in around 31 volumes), providing information 'not only about his own intellectual development and social network, but about Victorian science and society in general'.[4] And the *Victorian Lives and Letters Consortium*, coordinated by the University of South Carolina, has brought to light samples of life-writing from the period spanning the coronation of Queen Victoria to the outbreak of World War 1.[5] Documents include letters by Thomas Carlyle and the diaries of John Ruskin. Smaller scale projects include Sairio's (2009) study of correspondence by Elizabeth Montagu (as part of her research on letters of the bluestocking network), and Tieken-Boon van Ostade's (2011) study of unpublished correspondence of Robert Lowth. Finally, the *Network of Eighteenth Century English Texts* (NEET) corpus, developed by Fitzmaurice (2007), contains letters, fiction, prose drama and essays produced by Joseph Addison and members of his social milieu.

[2] University of Oxford (2009) *Cultures of Knowledge*. Available from: http://www.culturesofknowledge.org [Accessed 1 July 2015]. The *Cultures of Knowledge* project uses 'digital methods to reassemble and interpret…correspondence networks', using correspondence from the *Early Modern Letters Online* (EMLO) catalogue (dating from 1550 to 1750).

[3] Stanford University (2013) *Mapping the Republic of Letters*. Available from: http://republicofletters.stanford.edu/index.html [Accessed 1 July 2015].

[4] Cambridge University (2015) *The Darwin Project*. Available from: https://www.darwinproject.ac.uk/darwins-letters [Accessed 1 July 2015].

[5] University of South Carolina (2011) *Victorian Lives and Letters Consortium*. Available from: http://tundra.csd.sc.edu/vllc/ [Accessed 1 July 2015].

Other projects have incorporated letters by authors from a range of social backgrounds. *ARCHER: A Representative Corpus of Historical English Registers*[6] – initially constructed in the 1990s by Biber and Finegan (see Biber et. al. 1994a; 1994b) – for example, is a 'multi-genre historical corpus of British and American English covering the period 1600–1999'.[7] It is now managed by a consortium of participants, coordinated from the University of Manchester and contains – within the written registers component of the corpus – letters journals and diaries by authors from different layers of society. And the *Electronic Enlightenment* project at the University of Oxford has created an online collection of 67,875 letters and documents from the early modern period, including a 'myriad [of] unknown and ignored figures' including not only 'thinkers and scholars, politicians and diplomats, but also butchers and housewives, servants and shopkeepers'.[8]

However, for the most part, the correspondence projects described here contain letters by eminent persons and '[w]hile the available studies of educated letter writers certainly reveal interesting patterns of language variation within their correspondence [as well as revealing useful insights into a particular social group and historical context], the group of educated writers cannot be considered representative of the population at the time' (Auer and Fairman 2012, p. 78). Auer and Fairman argue that '[i]n fact, [in the Late Modern English period – 1700-1900] the educated only formed a small part of the population, as opposed

[6] University of Manchester (1990) *ARCHER: A Representative Corpus of Historical English Registers*. Available from: http://www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/ [Accessed 1 July 2015].
[7] For an overview of the project see: http://www.helsinki.fi/varieng/CoRD/corpora/ARCHER/updated%20version/introduction.html.
[8] University of Oxford (2008) *Electronic Englightenment*. Available from: http://www.e-enlightenment.com/info/about/ [Accessed 1 July 2015].

to the laboring poor who constituted between 60 and 70% ' (2012, p. 78). A similar trend can be observed elsewhere in Europe. While compulsory elementary schooling was only introduced in 1870 with the Elementary Education Act, by around 1800 many of the laboring poor could write at least something and as such formed the majority of those normally called 'literate' (Auer and Fairman 2012, p. 78; see also Cressy 1980). Yet their presence in historical corpora is noticeably lacking.

Since the introduction of Gutenberg's printing press in Europe in 1450 'European languages have been written in two modes: handwritten manuscripts and printed texts' (Fairman 2015 *forth.*). Fairman argues,

> …[o]f those two modes of writing linguists have almost always focused on printed texts. Therefore, not only have they written grammars and histories 'of the English language' from evidence which they draw from the mode with the smaller amount of written material, but they have written them from and about the variety used by the smaller part of the literate population: those who could write in or close to the print-worthy variety, now called 'Standard'…the variety which is defined according to the prevailing grammatical ideology' (2015 *forth.*).

Indeed, this situation has led several scholars to argue that the ideology which linguists use is biased.[9] Milroy, for example, argues that '…in English historical linguistics, this emphasis on elite language led to a situation in which only the

[9] While all ideology is 'bias' of one kind or another, the scholars referred to here, Milroy, Fairman et. al., are railing against the taking of the elite's language as a reflection of all kinds of English – and never taking 'uneducated' English as a good representation of English.

polite language of relatively educated speakers and writers could be included in historical descriptions of language. The language of the uneducated was not part of history: it could be ignored or rejected' (2012, p. 573), since the uneducated were/are not part of history (in the sense that history is made by those who write or control those who write).[10] Milroy goes on to explain that '[t]his ideologically-driven tendency to reject or ignore relevant evidence has been called the process of "erasure"' (ibid.), defined by Irvine and Gal (2000, p. 38) as:

> The process in which ideology, in simplifying the sociolinguistic field, renders some persons or activities (or sociolinguistic phenomena) invisible. Facts that are inconsistent with the ideological scheme either go unnoticed or get explained away. So, for example, a social group or a language may be imagined as homogeneous, its internal variation disregarded (cited in Milroy 2012: 575).

Furthermore, Milroy posits that one of the effects of using standard ideologies to understand historical texts is that it

> …project[s] the structure of present-day standard English on to past states of language, suggesting that these past states were largely invariant in structure, the researcher being free to use any argument that occurs to him/her to reject any evidence that does not fit in (2012, p. 581).

[10] Additionally, it could be argued that more sharply-felt social stratification (more contested, with a growing middle class and even educated working class) propelled the elite to strive to assert and maintain their superior distinctiveness by saying that a person is not really potentially of the elite unless you speak/write in this standard way.

By ignoring the different social groups and varieties of language, Milroy argues, we produce ‘an incomplete [or false] history of English’ (2012, p. 579), and, I would argue, society, since it is through language[11] that we are able to understand the lives and experiences of different social groups.[12] Milroy continues, ‘[t]he history that scholars…actually did give to the language was mainly a history of one variety – a relatively well-defined variety – the standard’ (ibid.).[13]

To summarise the discussion so far, certainly within the field of linguistics (and, arguably, other disciplines too) there has tended to be a focus on printed (rather than handwritten) manuscripts, typically produced by a relatively small number of people from the upper classes of society. This research has informed our understanding of language, reinforcing the notion of a ‘standard’ language, while everything else – without even looking at it – is ‘non-standard’. Looking at the broader picture, this process of prioritising printed material over handwritten material has led to a situation in which social groups, and the individual voices belonging to those groups, have been underrepresented, marginalised, or lost, thereby affecting our understanding of language and social history (issues that

[11] Not just through language (clothes, diet, religious practices, sports, courting practices, child-rearing, etc. also give useful insights into different social groups), but language is one good index.

[12] Here I am influenced by the work of Scott who argues that it is not possible to separate language and experience since language constructs identities, ‘position[s] subjects and produce[s]…experiences’ (1992, p. 25). The language used to talk about experience, therefore, not only reveals something about how a subject construes events and perceives the world, but it also reveals something about the discursive processes which helped to construct those experiences in the first place – a view which is certainly shared by linguists Halliday and Matthiessen (2004) and Hoey (2005). Whilst what underpins Halliday and Matthiessen’s work on systemic functional grammar is the notion of choice (the lexiogrammatical possibilities that ‘allow [a] speaker to represent the world in a particular way’ (Hunston 2006, p. 65)), what underpins Hoey’s theory of lexical priming is the idea that individuals are primed to use language in a certain way, therefore raising questions regarding the very notion of choice. Both theories, however, come from the standpoint that language and experience are inherently connected.

[13] On the issue of language standards and ideology see also, L. Milroy (1980), J. Milroy (2001), Street (1995), Mülhäusler (1996), Lass (1987), Greenbaum (1988) and Elspaβ (2007b).

have long been of concern in postcolonial and feminist theory).[14]

Arguably, a new, more inclusive, approach to looking at the history of a language is needed. For Watt, '"the" history of English, as it is presented in almost every introductory book on the subject, automatically leads novices in the field to the belief that a history of English is equivalent to a history of the standard language' (2012, p. 32) – Watt calls this the 'tunnel view'. An alternative way of conceptualising the history of language is, Watt argues, the 'funnel view' where

> the wide top of the funnel represents a period in the past in which there was no standard and in which we can find a number of linguistic varieties that seem to be related enough to be grouped together as 'a language'…as we move through time, the wide top of the funnel narrows to a neck through which language varieties must pass. The bottle would then be the container for the standard (2012, 586).

There are problems with the funnel model, which Watt himself points out: 'it displays a disregard for the historical trajectories of these varieties when a certain period of time is reached (in other words, when the narrow neck of the funnel is reached' (2012, p. 586); however, the funnel view at least acknowledges that there are different varieties of a language in the first place and that all of these varieties contribute to the resulting 'standard'. The question then is what should go into the funnel: whose voices have been captured so far and whose voices are missing? And what methodologies should be used to examine those voices?

[14] See studies by Loomba (1998), O'Hanlon (1988) and Spivak (1985; 2006) which discuss 'histories from below' (that is, from the perspective of the subaltern subject).

Fairman (2015) points to the growing interest in handwritten documents by lower-class writers by scholars in Germany (Vandenbussche 2006), Finland (Nordlund 2007), Russia (Yokoyama 2008), the Netherlands (van der Waal 2012) and other European countries. Fairman himself looks at poor relief letters by artisans and the laboring poor in England during the period 1750-1835 (2008; 2012). And there has been a growing interest in the nineteenth century emigrant letter too and how this type of material might inform our understanding of social history during a period of increasingly intense migration. Indeed, I believe that the emigrant letter is central to redressing the bias described by Fairman, Milroy and Watt, given the sheer amount of correspondence that crossed the Atlantic and Pacific oceans during the late eighteenth and nineteenth centuries in particular. The emigrant letter – the focus of this thesis – not only provides a better understanding, and a fuller picture, of how language was used, but, as will be discussed later, in the literature review, it also provides valuable insights into the migrant's experiences and context of writing. Especially interesting is the notion of outsider-becoming-insider status of emigrants, with regard to the newly-entered community and its dominant language. Cognitively, migrants negotiate edges – leaving Ireland and 'entering' America, for instance – and migrants, outsiders, travelers, anthropologists are often important reporters for seeing things that the indigenes have stopped noticing.

The sourcing, preservation and documentation of emigrant letter collections is now gathering pace, with the Internet providing a significant new forum for the dissemination of long-hidden archives.[15] Important studies of English (Gerber

[15] See, for example: The Mellon Centre for Migration Studies, Ulster American Folk Park Museum (2012-present) *The Irish Emigration Database (IED)*. Available from: http://www.dippam.ac.uk/ied/ [Accessed 1 April 2015].

2006), Scottish (Erickson 1972), Irish (Miller 1985; 2003; 2008), Welsh (Conway 1961) German (Kamphoefner et. al. 1988), Swedish (Barton 1990) and Norwegian (Zamper 1991) emigrants, among others, have demonstrated the value of using personal letters to gain a fuller and deeper understanding of both the complex social processes of migration and the conditions and daily lives of the emigrants themselves. They have also enriched our understanding of how the form of the letter itself, for emigrants more than any other group, 'functioned primarily to [reinforce and] reconfigure personal relationships made vulnerable by distance' (Elliott et. al. 2006, p. 17).

While it is clearly a good thing that more and more letter collections are being uncovered, documented, archived and analysed – offering a broader picture of language and society – the growing body of research is sometimes sporadic and disconnected. Researchers often work in isolation from one another and their projects evolve independently. Details of letter collections are sometimes difficult to find, and access to resources can be restricted by copyright and intellectual property concerns. This can lead to collections being missed or overlooked and/or the reduplication of work, with letters often being transcribed several times by different projects – projects which have their own research aims and, quite often, their own transcription and markup practices. Equally importantly, while different disciplines might use letter collections, the research is rarely interdisciplinary. The emigrant letter is often described as a site for multi-disciplinary research, but rarely is this put into practice. There seems to be a good case, therefore, for developing a collaborative, cross-disciplinary approach to

Immigration History Research Centre, University of Minnesota (2010-present) *Digitizing Immigrant Letters*. Available from: http://ihrc.umn.edu/research/dil/ [Accessed on 1 April 2015].

working with emigrant letter collections; the digital humanities, I would argue, offers a possible solution – bringing together scholars from across the disciplines to examine how technologies might be used to digitise, mark-up, search, visualise and analyse emigrant letters in ways that are useful to a range of users (academics and the general public) with a range of research interests.

Most existing digital letter collections consist of unannotated versions of original manuscripts. The digitisation process has made the letters more accessible and has also increased their searchability, at least to a certain extent. However, relatively few projects have moved beyond the digitisation stage to exploit text content and enhance usability and searchability through the use of digital technologies. Different emigrant letter collections cannot easily interconnect if they are simply digitised without markup, and some search pathways through the material will remain unavailable if software tools are not employed to process this encoding.

This thesis explores some of the opportunities and challenges of working with digitised historical emigrant letter collections. Essentially, the thesis does two things: first, it uses digital technologies (specifically corpus and computational methods of analysis)[16] to explore the language of emigrant letters, building on the existing body of research – primarily by historians – to offer another way into emigrant letters; second, it proposes a system of markup for capturing metadata relating to emigrant letters – metadata which can then be used to interconnect resources enabling users to carry out more nuanced and

[16] A corpus can be defined as a 'bod[y] of naturally occurring language data stored on computers' and corpus techniques of analysis as the 'computational procedures which manipulate this data in various ways . . . to uncover linguistic patterns which can enable us to make sense of the ways that language is used' (Baker 2006, p. 1).

sophisticated searchers.[17] I argue that my proposed system could be widely applied to emigrant letter collections, facilitating much greater interdisciplinary and collaborative analysis of such material than has been undertaken hitherto.

I will be working with one letter collection in particular: the Lough family correspondence (discussed in more detail in chapter one). Briefly, the Lough collection contains 99 letters by four sisters (Elizabeth, Alice, Annie and Julia) who emigrated from Ireland to America in the 1870s and 1880s. Although this thesis examines just one relatively small sample of letters, the methods used can be applied across collections. Chapters two to four use computational methods to examine the content of the Lough letters, while in chapters five and six I will demonstrate how an extensively TEI marked-up letter collection (or corpus) can be revealing and useful to discourse- or document-minded social historians and historical sociolinguists in ways that a 'bare' corpus would not be. Specifically, in chapter five I demonstrate how metadata relating to the document itself (its transcription history, provenance etc.) might be captured in a formalised way using TEI markup, and in chapter six I focus on metadata relating to people and places.

In uncovering, preserving and studying the emigrant letter it is possible to begin to redress the bias identified by Milroy et. al. Handwritten letters, such as emigrant correspondence, by authors from a range of socioeconomic and cultural backgrounds, contribute to a fuller, more complete history of language and social history, arguably helping us to move away from unhelpful notions of 'standard' and 'non-standard'. With advances in digital technologies it is now possible to

[17] Additionally, once digitised and encoded emigrant letters become multi-functional allowing 'several spin off products' to be 'extracted and realized, such as scholarly editions, reading texts, indexes, catalogues, calendars, regests, polyfunctional research corpora etc.' (Vanhoutte and Van den Branden 2009, p. 94).

document and search this growing body of material in new and creative ways and there is a real opportunity, through knowledge transfer and data sharing, for genuinely collaborative, cross-disciplinary research between programmers, computational linguists, corpus linguists, historians, migration studies scholars and archivists.

LITERATURE REVIEW

Personal and corporate letters of eminent persons have long been used for social, historical and cultural studies. Such letters are saved, transcribed, edited and published. Eminent persons themselves are sometimes part of this process – making copies of their own letters before sending them, or retrieving letters from recipients. The practices of Thomas Jefferson, for example, have received much attention: his efforts to repair the research record and his use of the letterpress and polygraph letter duplication systems (see Sifton 1977). Over the past decades, however, there has been a growing interest in what scholars have termed 'history from below' or 'intrahistoria'[18] – that is, a history of the popular classes. Lyons makes a useful distinction between what he terms the 'old history from below' and the 'new history from below'.[19] The 'old history from below', Lyons argues, sought to '…incorporate the lower classes into the general historical narrative through "number and quantity"' (Lyons 2010, p. 14 citing Kaye 1984, p. 225). As a result, 'the subordinate classes remained a silent and disincarnated mass without any personal identity' (Lyons 2010, p. 14). The 'new history from below', on the other hand, is more 'individualised' and 'sensitive to the voices of

[18] A term coined by the Spanish writer Miguel de Unamuno in 1985. 'It refers to the value of the humble and anonymous lives experienced by ordinary men and women in everyday contexts which form the essence of normal social interactions, as opposed to the lives of leaders and famous people that are generally accounted for in canonical histories' (Amador-Moreno et. al. 2015 *forth.*).

[19] In discussing the 'old history from below' Lyons refers, in particular, to the Annales School. Founded by Marc Bloch (1886-1944) and Lucien Febvre (1878-1956), the Annales School 'promoted a new form of history, replacing the study of leaders with the lives of ordinary people and replacing examination of politics, diplomacy, and wars with inquiries into climate, demography, agriculture, commerce, technology, transportation, and communication, as well as social groups and mentalities' (*Encyclopedia Britannica*. Available from: http://global.britannica.com/topic/Annales-school. [Accessed 30 June 2015]). Lyons attributes the term 'new history from below' to Hitchcock's 2004 review of Sokoll's *Essex Pauper Letters* (2001).

the poor' (Lyons 2010, p. 16). In summary, while the 'old history from below' is a history which 'remained collective and largely impersonal' (Lyons 2010, p. 15), the 'new history from below' is a history which foregrounds the perspective of ordinary individuals who lived and experienced historical, social, economic and cultural change and its various consequences. Lyons (2010, p. 16) suggests that the 'new history from below' is 'new' for the following four reasons:

1. It re-evaluates individual experience.

   (The 'new history from below' is a history on a 'micro-historical scale' (Lyons 2010, p. 17).)
2. It searches for the personal and private voices of *la gente commune*, however they may be mediated through institutional or other channels.

   (As already mentioned, until relatively recently only the writing of educated, eminent persons attracted the attention of cultural historians. The 'new history from below' seeks to understand the experiences of the 'poor and uneducated' through examining the writings of the poor themselves (Lyons 2010, p. 19).)
3. It modifies the direction taken by the linguistic turn against which it is in some sense a reaction.

   (Lyons explains that while 'protagonists of the "linguistic turn"[20] were intent on deconstructing dominant discourses' they 'neglected to

[20] Lyons describes the 'linguist turn' as follows: 'Post-modernist influences on historiography which we loosely call the "linguistic turn", have forced us to re-consider the history we write. We now recognise that the texts we ourselves compose obey certain rules and conventions and adopt certain strategies. Our own history-writing has a literary aspect, in the sense that it constructs a narrative and deploys certain rhetorical strategies to persuade the reader. The history we write is never a transparent account; it is a text, an artifact which refers to other texts, can only be understood in conjunction with other texts and uses literary devices to sway and convince' (2010, p. 20).

consider how such dominant discourses were actually consumed' (2010, p. 20). The 'new history from below' directs 'the techniques of deconstruction and discourse analysis towards the texts…of the subordinate classes' (ibid.). Through this process it may be possible to better understand 'the lower-class assimilation of national myths, languages and beliefs' (2010, p. 21).)

4. It considers ordinary readers and writers as active agents in the shaping of their own lives and cultures.

   (The 'new history from below' 'recognises the autonomy of lower-class writers (and readers), and refuses to regard them as passive receptacles for information and ideologies produced by someone else' (Lyons 2010, p. 21). Lyons refers to dominant discourses surrounding the subject of migration arguing that '[t]he problem with the socio-economic approach is that it treats emigrants as people responding passively to impersonal changes like industrialisation and fluctuations in the labour market. It deprives them of any independent choice' (2010, p. 21).)[21]

The personal letter has been one of the main mechanisms for accessing and understanding 'history from below'. While many important studies in British and European history have focused on rescuing the voices of the poor and retrieving 'working-class "ego-documents" and autobiographical writings' (Richards 2006, p. 58) – see for instance the collection of *Essex Pauper Letters, 1731-1837*

[21] Lyons refers, in particular, to Martínez's 2006 study of the Moldes family correspondence between the Asturias and Chile, which reveals the 'complex relationship between individual choice and family strategy' in the context of migration (2010, p. 21).

written by, or on behalf of, paupers seeking support from the local poor law in the county of Essex[22] – there has been little 'sign of any convergence with recent comparable work on emigrant letters' (Richards 2006, p. 59) despite the fact that migrant correspondence frequently does reach into comparable layers of society.[23] At the same time, however, the emigrant letter is different from the pauper letter in that 'it was rarely the plea of "the powerless to the powerful".[24] Emigrants were likely to have been much more literate than paupers (though some were both paupers and emigrants)' (Richards 2006, p. 61).

For some scholars, the emigrant letter provides 'the unmediated voice…the voice of pure experience' (Elliott et. al. 2006, p. 7) – O'Farrell, for example, examining Ulster emigrants to Australia, maintains that correspondence provides 'an intimate insight into what the migrant actually thought and felt, expressed without constraint, and with the honesty and candour appropriate to close family situations' (O'Farrell 1984, p. 3, cited in Fitzpatrick 2004, p. 25). However, as Elliott et. al. point out, this is not entirely true as writers were almost certainly influenced by the language of church or politics and 'most probably they learned to write letters by reading the letters of others' (2006, p. 7).[25] Nevertheless, whilst recognising the influence of the context of situation (the circumstances in which the letter was produced) or broader still the context of culture (the societal pressure for the author to perform in a particular way – by writing the letter in the first place and by respecting a particular culture of letter writing when doing so),

[22] Other studies which focus on the writings of the poor include Burnett et. al. (1984), Fairman (2000), Lorenzen-Schmidt and Poulsen (2002) and Yokoyama (2008).

[23] Although emigrants came from a range of socioeconomic backgrounds the vast majority were, as Erickson puts is, 'ordinary working people' (1972, p. 1).

[24] Here Richard's is using James Scott's phrase, quoted in Hitchcock, King, and Sharp (1996, p. 6).

[25] In other words, 'immigrant writers were immersed in cultures which informed their often tentative writing' (Elliott et. al. 2006, p. 7).

I would argue that nineteenth century emigrant correspondence gets as close as possible to the letter writers' lived experience. The reality of the authors' lives – or the way in which the authors construed their experiences – is revealed through their writing.

As Helbich and Kamphoefner quite rightly point out, of the millions of emigrant letters sent and received 'only a tiny, infinitesimal fraction has been preserved and is available to researchers' (2006, p. 29). However the numbers are still significant and, as suggest by Fay, '…if we had but one per cent [of emigrant correspondence] before us, we could attempt a real history of emigration' (Fay 1951, p. 262, cited in Richards 2006, p. 56). This 'infinitesimal fraction' – thousands of letters exchanged between emigrants and family and friends in the homeland – provides scholars with a unique source material with which to explore and understand the individual emigrant experience as well as nineteenth and twentieth century mass migration to the Americas, Africa and Australasia more generally. However, the plenitude of research material brings with it methodological challenges. As Lyons puts it, 'the problem with ordinary writings is not that they are scarce and ephemeral, but that there is such an abundance of them that the historian hardly knows where to begin' (2010, p. 16). Additionally, 'balancing the individual (or the small group) against the broader history of social and cultural change' (Lyons 2010, p. 18) is an ongoing challenge. This problem of 'reconciling the individual with the general' (Lyons 2010, p. 14) is echoed elsewhere (see, for example, Richards (2010; 2013)); I will return to this issue later in the review.

For Gerber, emigrant letters have generally been used in one of two ways: to 'provide color and drama in historical narratives, or to document societal-level and group-level generalizations', or as edited collections which 'let the letter-writers speak for themselves, while providing some background information that enables readers to place the [author] in the general societal framework of a certain place and time' (2006, p. 31). In the study of the emigrant letter a good starting point is the work of Thomas and Znaniecki (1918; 1919-1920), whose research on Polish migration to America examined, amongst other things, 'how social solidarity was maintained within families among Polish peasants both in Poland and in the United States' (Elliott et. al. 2006, p. 5). Other influential studies – Erickson (1972) (examining English and Scottish migration), Kamphoefner et. al. (1988) (German migration), Blegen (1955) (Norwegian migration), Conway (1961) (Welsh migration), and Barton (1975) (Swedish letters), to name just a few – have demonstrated the value in using personal letters to gain a fuller, multi-perspectival understanding of both the complex social processes of emigration (such as push/pull factors, and the role of family, communities and institutions) and the conditions and daily lives of the emigrants themselves. Studies examining Irish migration have contributed significantly to the growing interest in emigrant letters as a primary data source. In the 1950s, Arnold Schrier[26] formed an alliance with the Irish Folklore Commission to harness the Commission's existing research methodologies – questionnaires, tape recorders, and a network of interviewers – to collect information about Irish emigrant letters (Miller 2008, p. xii), and in 1955 he broadcast – in newspapers and on the radio – appeals to the Irish people to donate any emigrant letters they held in their possession (Schrier 1958). Some of these letters were later passed to

[26] Arnold Schrier, Professor Emeritus at the University of Cincinnati (http://www.artsci.uc.edu/faculty-staff/listing/last_name_alpha.html?eid=schriea&thecomp=uceprof).

Kerby Miller[27] who developed the use of Irish emigrant letters in important works including *Emigrants and Exiles: Ireland and the Irish Exodus to North America* (1985), in which he argues that 'Irish-American homesickness, alienation, and nationalism were rooted ultimately in a traditional Irish Catholic worldview which predisposed Irish emigrants to perceive or at least justify themselves not as voluntary, ambitious emigrants but as involuntary, nonresponsible "exiles," compelled to leave home by forces beyond individual control, particularly by British and landlord oppression' (p. 556).[28] Miller's own archive of Irish emigrant correspondence – gathered over many years – now exceeds five thousand letters together with extensive background information relating to each individual author.[29] Fitzpatrick (1994), using a much smaller sample of letters (seventeen sequences of letters between 1843 to 1906), examines Irish migration to Australia. Taking more of a discourse analytic approach, Fitzpatrick finds no comparable 'exile' trope. Instead, he observes how the letters are 'a tool for sustaining solidarity among separated kinsfolk and asserting individual rights within family and neighbourhood networks' (1994, p. 35). I will be referring to the work of Schrier, Miller and Fitzpatrick throughout this thesis – most notably in chapters one and four.[30]

The studies outlined so far certainly show the value of using ego-documents to understand the migratory patterns and experiences of different

[27] Kerby Miller, Curators' Professor at the University of Missouri (http://history.missouri.edu/people/miller.html).
[28] This argument is challenged by McCaffrey in the preface to his 1992 book *Textures of Irish America*: first, Miller is attacked because he undermines the standard Irish-American 'rags to riches' story; second, he is attacked on methodological grounds.
[29] See also Miller (2003; 2008).
[30] Other studies which use emigrant letters to examine aspects of Irish migration include: McCarthy (2005) who adopts a similar approach to Fitzpatrick to examine Irish migration to New Zealand; O'Farrell (1984) who looks at Ulster emigrants to Australia and Atkinson's (1997; 2005; 2014) three volumes on European migration to Australia.

social and ethnic groups. On the surface, at least, these studies might appear to fall into the tradition of 'old history from below' in so much as they present a collective experience of migration: the Norwegians in North America, or the Irish in England, for example. Indeed, Ghaill and Haywood (2010), discussing Irish migration history, point out that 'representations of generations of emigrants have been subsumed under hegemonic images of post-Famine emigration with their overarching motif of exile' (2010, p. 385). And Gerber argues that while 'social historians have been especially skilled in understanding large categorical social groups – social classes, ethnic groups, religious denominations, and men and women' they have sometimes failed to understand 'the individual self, in relation to other individual selves, [in relation] to the world' (2006, p. 32).[31]

Ghaill and Haywood's own study examines 'concepts of home, nationality and belonging by evaluating explanations of (e)migration of mid-20th century Irish working class men' (2010, p. 385). They do this by carrying out semi-structured interviews with 24 Irish men aged between 54 and 76, who had emigrated from Ireland to England between 35 and 55 years previously. Ghaill and Haywood argue that 'contemporary research shares an incisive self-reflexivity that enables the disruption of established migration/diaspora research rationalities that serve to rigidly catalogue the lives of transnational migrant subjects' (2010, p. 386). For Ghaill and Haywood, 'it is important to disturb such rationalities and in effect challenge "settled" epistemological positions that permeate approaches to migration, diaspora and national belonging' (ibid.). The 'new history from below', which priorities the individual, has the potential, then,

[31] However, the fact that the conclusions of these studies are based on the writings of (hundreds, sometimes thousands of) individual emigrants would surely place them within the 'new history from below' category.

to reconfigure existing discourses and ways of knowing; challenge 'containing categories' that 'insist on the epistemological security and stability of the object of inquiry' (Ghaill and Haywood 2010, p. 387),[32] and question assumptions about 'home and nation, ethnic and racial (in)visibility and the imbrication of social and cultural identities' (ibid., p. 396).

Using letters as his data source, Gerber's study of British emigrants in America also focuses on the individual. Viewing the personal letter as an object of study in its own right, Gerber examines how emigrant correspondence embodies relationships, experiences and mental worlds (2006). More recent studies, taking a similar approach to Gerber, have examined how transatlantic relationships are changed and maintained, identities assimilated and narratives constructed and performed (see, for example, studies by Eyford (2015), De Fina and King (2011), Cancian (2010), DeHaan (2010) and Harper (2010)). In most of these studies, the researcher is inferring outwards from the letter, taking the content of the letter to then make claims about what that content means, or what it reveals about the context of situation and culture. This research certainly provides valuable insights into the individual emigrant's experience; however, methodologically speaking, the conclusions are potentially open to criticism firstly because they offer just one interpretation of an individual's letters and secondly because there are no explicit means of replicating, testing, or building on the findings.

Recent research in socio-, historical- and corpus-linguistics perhaps goes some way to addressing this methodological issue. Echoing the arguments raised by Milroy (2012), outlined in the Introduction to this thesis, Elspaβ argues that

[32] Here, Ghaill and Haywood refer to the work of Massey (2005).

‘most language histories of the Western hemisphere tell the story of *printed* languages’. He goes on to say that ‘a complete account of language history, viewed from the perspective of its agents, can only be achieved if we attempt to consider as many text sources from as many different times, varieties, regions, domains, and text types as possible’ (2012, p. 156). Furthermore, Elspaβ argues that ‘having to depend on written documents…does not mean that it is impossible to study the history of speech from such sources’ (2012, p. 157):

> The traditional distinction between “spoken language” and “written language” is simplistic and even misleading. To arrive at an adequate understanding of the nature of “speech,” “spoken language,” and/or “orality,” it is essential to place these notions into an integral model (ibid.).

The model that Elspaβ suggests is Koch and Oesterreicher’s (1985; 1994) notion of ‘language of immediacy’ (for example, an intimate conversation) versus ‘language of distance’ (for example, a written legal contract), with all text types being positioned somewhere in between. Thus, the personal letter, leaning more towards the notion of ‘immediacy’ because of its intimate nature, might provide useful insights into language change and variation with regard to both speech and writing. Additionally, personal correspondence, ‘rather than just supplementing existing language histories with some aspects of orality[,]…can serve as text sources fundamental to a radically different approach to language history in its own right’ (Elspaβ 2012, p. 160) – an approach that is described by Elspaβ and others as ‘language history from below’.[33] This approach is an attempt to write ‘alternative histories’ (Watts and Trudgill 2002 cited in Elspaβ 2012, p. 161) to counterweight ‘the tendency of traditional language historiography to ignore or trivialize the histories of minority languages, varieties, and registers that did not win the race, and to ignore or shrug off linguistic variation and digressions from the dominating varieties as corrupted language, and therefore non-valuable data for linguistic research’ (Elspaβ 2012, p. 161).

Montgomery’s (1995) study of Ulster Scots was one of the first to demonstrate how emigrant correspondence provides access to the language of the “common” people. (Subsequent studies include Cano Aguilar (1996), García-Bermejo Giner and Montgomery (1997), López Álvarez (2000) and Elspaβ (2007a).) These studies adopt a bottom-up, empirical approach to studying emigrant letters, taking as their starting point words and phrases, and then looking at how these words and phrases typically behave in sentences, paragraphs and texts, before considering what these linguistic patterns or phraseology[34] might reveal about the situational and cultural contexts in which the letters were produced. Dossena (2007), for example, examines the use of formulaic as well as dialectal features of language in a corpus of nineteenth-century Scottish emigrant letters to see how such linguistic strategies contribute to, and reinforce, social

[33] See also Elspaβ (2007b) and Elspaβ et. al. (2007).

[34] The term ‘phraseology’, in this thesis, refers to the way in which a word typically behaves in context – both the grammatical position it tends to adopt and the words it tends to collocate with. Phraseology and collocation are linked, but whereas collocation tends to refer to word pairings, phraseology refers to extended patterns, where meaning might be carried over several words. The regularities (or patterning) and subtleties in the usage of a word, Hunston argues, are ‘difficult to intuit, and [are] observable only when a lot of evidence is seen together’ (2002, p. 12). Idioms are fixed expressions and linked to metaphor and figures of speech, but could also be described as phraseology - whereas idioms tend to be fixed and self-contained, phraseology is more open to subtle variations.

bonds between author and recipient.[35] (In its most general sense, a corpus is a collection of texts, designed to be representative of the way that language is used in a particular context.) To do this, Dossena looks at a subsection of letters from the *Corpus of Nineteenth-Century Scottish Correspondence*: forty-two letters (approximately 27,000 words), dating from 1815 to 1892, by thirteen male informants and two female informants.[36] A close qualitative study of the linguistic features characteristic of the letters teases out some interesting findings. Dossena observes that 'involvement strategies' in the openings and closings of the letters were 'mainly dependent on the conventions of formulaic usage' as set out in letter writing manuals of the time. However, she goes on to say that 'within the body of the letter . . . encoders express their psychological proximity to their recipients by means of other linguistic devices', such as the use of Scotticisms (dialect, which, as observed by Dossena, is often employed humorously to stress a 'common cultural background'); visualisations of context (descriptions of people, places and likenesses) and epistemic modality (words which express certainty/probability, such as *might* and *suppose*, which are used to 'predict the recipient's reactions or the encoder's suppositions about what is going on at home') (2007, p. 21). Although empirical in nature and taking a more bottom-up approach to identifying salient linguistic features across a range of texts, this study is still primarily qualitative and therefore open to the same criticisms previously mentioned. The conclusions resulting from (what are very interesting) observations would have greater strength if it were possible to test their significance reliably. Are these observations typical, unusual, or evenly distributed across different authors, for example? Quantitative investigation

[35] See also Dossena (2010).

[36] For full details of the design and contents of the corpus see Dossena (2004).

and/or statistical tests would help to make claims about the relevance of these observations. Arguably, without quantitative support, it is difficult to appreciate fully the significance of the linguistic features being noticed.

To demonstrate the value in applying statistical measures to test, challenge or support qualitative observations, McLelland (2007) used quantitative methods of analysis to examine the language of nineteenth-century German emigrant men's and women's private correspondence. Referring to twentieth-century studies in language and gender,[37] McLelland argues that research into gender differences has lacked clarity, in part, because it has been qualitative rather than quantitative: '[p]roblems arise when data that are essentially anecdotal in nature are treated as if indicative of general trends without appropriate statistical analysis' (2007, p. 46). In her study, McLelland focuses on some of the linguistic strategies identified in recent scholarship as being more typical of women in conversation – such as the use of epistemic modality (as previously mentioned, words like *might* and *suppose*), hedging devices (words like *seem*, *believe* and *sometimes*), and question tags (such as *isn't it? shouldn't I?* and *don't they?*) – and then uses statistical methods to test whether such gender differences are evident in a corpus of nineteenth-century letters. The analysis involved using two corpora: Corpus One (a pilot corpus) containing twenty-two letters by women and twenty-two by men (approximately 30,000 words), dating from 1850 to 1900 and representing seven female and eight male authors; Corpus Two (a much larger, more representative corpus) containing ninety-one letters by men and ninety-one by women (approximately 84,000 words), from the same time period

[37] See, for example, Bergvall and Freed (1996), Cheshire and Trudgill (1998), Holmes (1995), Kotthoff and Wodak (1997), Talbot (1998), and Wodak (1997).

and representing thirty-eight female and thirty-eight male authors. Of the linguistic features investigated in Corpus One, only the discourse particle *doch* (often used as an intensifier or emphatic device) showed any significant difference between genders, being more frequently used by female authors in phrases such as *ich denke* (broadly translated as *I think*) to soften assertions. The data also gestured towards female authors being more likely to 'soften imperatives, express more wishes, and [be] more emphatic in their formulations than . . . men', although these findings were somewhat tentative (2007, p. 55). However, when the same investigation was carried out using the larger corpus (Corpus Two), these findings received little statistical support. What the findings did show, however, was that the female authors used more intensifying adverbs (in English these would be words like *very*, *really* and *so*), they were more likely to address the recipient in the body of the letter and they referred to themselves using the first person (*I*) more frequently than their male counterparts. The data also showed that the female authors tended to adopt more politeness strategies – *bitte(n)* (a verb meaning to ask/request) and *bitte* (similar to *please* or *you're welcome*) when making requests – however, as McLelland points out, this finding could simply be a result of more requests being made by women than by men in the first place. McLelland hypothesises that the high frequency of *doch* in Corpus One may be explained in terms of educational background. It is one particular author who contributes over a third of all occurrences of *doch* in Corpus One and this author also adopts a more colloquial, speech-like style in her letters, indicative of a lower level of education.

Chapters two and three of this thesis will look in more detail at some of the linguistic studies that have been carried out with regard to correspondence.

However, my point here is to highlight how McLelland's study (and other quantitative studies by, for example, Elspaβ (2002) and Brinks (1991))[38] demonstrates the possibilities and opportunities of using more quantitative methods of analysis to explore emigrant letters – methods that can be tested, verified and built upon. Additionally, quantitative work on non-emigrant letters may also provide possible avenues for future research. Researchers at VARIENG[39] (the Research Unit for the Study of Variation, Contacts and Change in English), at the University of Helsinki, use (quantitative) corpus and computational methods of analysis to examine, amongst other things, language change and variation in seventeenth and eighteenth century personal correspondence. As will be discussed in chapter four, the emigrant letter poses various challenges in terms of using automated corpus and computational tools, as quite often the letters lack punctuation and contain irregular spellings and grammatical constructions. Nevertheless, the same methodologies and approaches that are being used with non-emigrant letters could certainly be trialed with emigrant correspondence and tools such as VARD[40] (a pre-processing tool designed to deal with spelling variations in historical texts) may help with this process. There is certainly a lot of potential for using quantitative methods of analysis with historical emigrant letter collections; it is, however, a very difficult balance to achieve between offering a rigorous, replicable and

[38] Elspaβ (2002) examines language change and variation using the writings of 'ordinary [German] people'; Brinks (1991) examines how the notion of 'old world' and 'new world' is textually performed in several Dutch American letter series.

[39] *VARIENG*. University of Helsinki. Available from: http://www.helsinki.fi/varieng/index.html [Accessed 26 November 2015].

[40] *VARD*. UCREL (University Centre for Computer Corpus Research on Language), Lancaster University. Available from: http://ucrel.lancs.ac.uk/vard/about/ [Accessed 26 November 2015]. See, also, Baron, A. and Rayson, P. (2008). VARD 2: A tool for dealing with spelling variation in historical corpora. Proceedings of the Postgraduate Conference in Corpus Linguistics, Aston University, Birmingham, UK, 22 May 2008.

systematic quantitative approach whilst at the same time not losing sight of the very personal, idiosyncratic and subjective data at hand.

To summarise the discussion so far, the research outlined above clearly demonstrates the value in using emigrant letters as a primary data source. As Richards (2010) puts it: 'Emigrant letters speak for the individual letter-writer but, in sufficient numbers, they also create a collective account of the world into which they were relocated, uprooted or otherwise' (pp. 3-4). While some studies use emigrant letters as a way into understanding the collective – 'historians commonly extrapolate informally and unconsciously from individual testimony towards a view of "the spirit of the age", or the common "mentalities", or "ways of thinking" of the times' (Richards 2010, p. 4) – other studies are more interested in 'the dense specificity of personal experience' which, Lodge argues, 'is always unique, because each of us has a slightly or very different personal history, modifying every new experience we have' (Richards 2010, p. 3 citing Lodge 2002, pp.10-11).

However there are problems: studies that prioritise the collective might be accused of assigning the emigrant to an anonymous mass, thereby silencing the various individual voices, while studies that prioritise the individual might be accused of not doing anything more than offering an individual's biography and life story. As Richards points out, 'reassert[ing] the traditional use of emigrants' letters, returning them to the micro-historical form' simultaneously 'denies that emigrants' letters can be made the basis of any kind of historical sociology' (2010, p. 13).[41] Richards goes on to say that this approach (which focuses on the

[41] Here, Richards is referring to Gerber (2006) who argues that 'emigrant letters, or indeed any type of personal correspondence, is almost always a commentary on the individual psyche of the writer…[letters] are restricted to the way individual writers recreated their own personalities, their

individual psyche of the writer) ‘is an austere and severely constraining formulation which restricts the source too much, and empties out the baby with the bath water’ (ibid.). And while quantitative methods of analysis would complement many of the more qualitative studies outlined so far – offering a replicable evidence-based approach that allows researchers to test hypotheses[42] – they, too, have their limitations. Quantitative studies can lose sight of the individual voices (the emigrants behind the statistics), and the findings of such studies arguably have limited use if they are not viewed within their social, historical and cultural context – that is to say, within a multi-disciplinary framework. As O’Sullivan argues, ‘no one academic discipline is going to tell us everything we want to know about the Irish [or other] Diaspora. The study of migration, emigration, immigration, population movements, flight, scattering, networks, transnational communities, diaspora – this study demands an interdisciplinary approach’ (O’Sullivan 2003, p. 131).[43] Finally, there is the issue of representativeness among emigrant letters. Richards points out that for some scholars ‘the writing and survival of emigrant letters is haphazard and highly selective and…too small for the purpose of representing large phenomena’ (2010, p. 13). Indeed, most studies that use emigrant letters as their source begin by discussing the problem of dealing with skewed representativeness. Erickson (1972), for example, writes of ‘the paucity of letters of laboring immigrants’, observing that

emotional conditioning to the experience of emigration, reformulating their relationships and reconstructing their personal identities’ (Richards 2010, p. 13 summarising Gerber 2006).

[42] Richards (2010), for example, references the work of Belich who ‘posits “a remarkable shift in attitudes”, that is, an attitudinal change which actively propelled…[the migration] of millions of British people with massive consequences across the globe’ (p. 10); Richards argues that ‘we need firmer evidence and also ways of testing the psychological transition upon which his hypothesis is founded’ (ibid.). Quantitative and corpus methods of text/content analysis may provide the sort of evidence Richards is referring to.

[43] See also Brettell and Hollifield (2000, p. vii).

> any sample of immigrant letters gives undue emphasis to people who failed as immigrants, and those who did not break their ties with the homeland: the poorest emigrants, those who emigrated for the most straightforward economic reasons, will not be found among the letter-writers (Richards 2006, p. 58 citing Erickson 1972, p. 7).

Yet, as Richards points out, 'historians always deal in fragments of evidence' (ibid.) and as increasingly more letter collections are uncovered, so our understanding of the emigrant experience will broaden and evolve, if, that is, we adopt new technologies to categorise and interpret the huge quantity of surviving material.

The overarching question, then, is not so much about whether we focus on the individual or the collective (so long as we are looking at the writings of the emigrants themselves), it is more to do with how we reconcile the individual with the general; the fragments with the whole (Lyons 2010, p. 14). In other words, how do we study individuals against the broader historical context and in so doing 'give a human dimension to significant historical issues'? (Lyons 2010, p. 18). The solution, I want to argue, lies in the digital humanities (which includes corpus and computational linguistics). A multi- and inter-disciplinary field by definition, the digital humanities offers the opportunity for bringing together scholars from across the disciplines to look at ways of harnessing the power of new technologies in the study of historical emigrant letters. Once letters are digitised and annotated in a formalised and consistent way it is possible to interconnect collections, allowing the user to constantly move between the individual and the whole – comparing individual letters against letter series, noticing what is unique or different as well as patterns and trends.

Alliances between disciplines are starting to form. These alliances are largely driven by the need for reliable digital transcriptions of manuscripts, which can be used across a range of disciplines. The *Digital Editions for Corpus Linguistics* (DECL) project, for example, 'aims to create a framework for producing online editions of historical manuscripts suitable for both corpus linguistic and historical research' (Honkapohja et. al. 2009, p. 451). DECL argue that 'up to now, few digital editions of historical texts have been designed with corpus linguistics in mind. Equally, few historical corpora have been complied from original manuscripts' (ibid.). They go on to say that 'most of the problems associated with using traditional historical corpora stem from the fact that…the transcription and digitisation of original manuscript texts into machine-readable form takes a lot of time and expertise' (Honkapohja et. al. 2009, p. 456). 'Most historical corpora', they are argue, 'are based on printed editions, which "have generally not been produced with linguistic study in mind, and may not always be reliable"' (ibid. citing Kyotö et. al. 2007, section 3). For example (as will be discussed in chapter five), editions can often 'compound multiple manuscript witnesses into a single text' and apply 'varying editorial principles' (Honkapohja 2009, p. 456), making it difficult – if not impossible – to access the original language of the document.[44] Additionally, 'textual editors tend to focus on texts considered culturally or literarily "significant", and relying solely on editions can lead to the omission of whole categories of material' (Honkapohja 2009, p. 457)

[44] Elspaβ, for instance, observes that 'for limitations of space, many printed editions of letters and diaries tend to omit lengthy formulae, quotes, and other prefabricated language. Such language material may also appear tedious for a reader who is interested in a content analysis. Linguistically, however, this sort of information is highly telling' (2012, p. 164), potentially revealing something about 'verbal rituals of a period' or 'the extent to which a writer was familiar with the conventions and fashions of formal letter-writing of that period' (ibid.).

– exactly the sort of situation that scholars interested in ‘history from below’ are trying to avoid. As Nurmi puts it, there was a tendency amongst nineteenth-century editors ‘to edit only the letters of historically important people, and ones describing important historical events. Editors often disregarded family letters concerning everyday life’ (Nurmi 1999, p. 54 cited in Honkapohja 2009, p. 457). There are also issues relating to the integrity of printed editions. As Honkapohja et. al. point out, ‘few pre-1980s editions provide detailed information about their practices concerning orthography and frequently normalise spelling – not to mention punctuation – to varying degrees’ (2009, p. 458).[45] Copyright also poses problems: although historical documents (from the Medieval and Early Modern periods) are typically free of copyright, ‘modern printed editions of these documents usually are not’ (ibid.). And, finally, using digital editions for corpus compilation involves a certain amount of manpower. The compiler will either need to transcribe the printed edition or use Optical Character Recognition (OCR) technology to scan it, which, as will be discussed in chapter one, comes with its own unique set of problems. In both cases, however, at least some degree of proofreading is required meaning that new errors are likely to be introduced into the text (Honkapohja 2009, p. 459). And there are problems in the way that historical corpora are compiled too. Many projects use project-specific encoding practices, often borrowed from earlier projects and adapted for their specific requirements. This, in turn, ‘limits the development and use of common tools and the convertibility of corpora from one format to another’ (Honkapohja et. al.

[45] Elspaβ, for instance, reports that ‘there is some dispute as to whether manuscripts should be presented as “quasi-facsimiles” or not’ (2012, p. 165). He explains that while ‘Hunter (2009, pp. 72-85) argues that it is not necessary to reproduce, for example, ligatures or tildes to denote a duplication of n or m and makes a case for even expanding abbreviations, such as Mtie to Majestie [other] linguists…consider such interventions as too far-reaching’ (ibid.).

2009, p. 459-460). This, Honkapohja et. al. argue, is surprising given that the Text Encoding Initiative (TEI), providing a standard for the electronic encoding of textual data, have been around since the 1990s (2009, p.460). Additionally, much more information could be annotated when compiling historical corpora – not just linguistic information. If, for example, information about the materiality or structure of the manuscript were captured, this 'deeper representation' would help 'to widen the applicability of the corpus to different types of research' (Honkapohja et. al. 2009, p. 460).

There are, it would seem, lessons to be learnt from both sides: manuscript studies and corpus linguistics. In developing the DELC framework, Honkapohja et. al. call for faithful representations of manuscripts (with transcription practices being fully documented) in machine-readable and fully searchable format, using standoff annotation to mark-up linguistic and non-linguistic information, thus allowing (small) corpora to expand and interconnect on various levels. This thesis responds to several of the issues raised by Honkapohja et. al., with a special focus on emigrant letters. In particular, chapters five and six propose a system of markup for capturing metadata relating to emigrant correspondence – some of this metadata is gathered by using corpus and computational methods to analyse the letters themselves (as demonstrated in chapters one to four).

The field of digital humanities has experienced significant growth in the past few years, especially with regard to the development and creation of markup standards. As Honkapohja et. al. point out,

> perhaps the most significant effort in providing standard forms of textual markup has been the Text Encoding Initiative (TEI).[46] Currently in their fifth public version (P5), the XML-based *TEI Guidelines* have been adopted by a large number of projects within the field of digital humanities, including the *British National Corpus* (BNC)[47] and even some historical corpus projects, such as the *Corpus of Northern English Texts from Old to Early Modern English* at the University of Seville[48] (2009, p. 465).

In terms of correspondence projects that are adopting TEI, two of the most notable are the *Digital Archive of Letters in Flanders* (DALF) and the *Carl Maria von Web Collected Works* (WeGA).[49]

The DALF project was perhaps the first of its kind to propose a formal TEI-customised framework for annotating correspondence (Vanhoutte and Van den Branden 2009, p. 77) – in this case 1,500 letters by Dutch (and sometimes French) speaking writers, including authors, composers, musicians, critics, illustrators and publishers, together with their family, friends and colleagues.[50] This framework was later developed by the WeGA project, which explored and presented the correspondence of the composer Carl Maria von Weber (1786-1826), whose letters will be published first in digital format on the website and

[46] TEI Consortium (eds.), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Available from: http://www.tei-c.org/Guidelines/P5/ [Accessed 1 August 2015].

[47] *The British National Corpus*, version 3 (BNC XML Edition) (2007) Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Accessed from: http://www.natcorp.ox.ac.uk/ [Accessed 1 August 2015].

[48] For a list of articles and papers that have used the corpus in their research see: http://www.helsinki.fi/varieng/CoRD/corpora/SCONE/bibliography.html.

[49] *Carl Maria von Weber – Collected Works* (WeGA). Universität Paderborn. Available from: http://www.weber-gesamtausgabe.de/en/Index [Accessed 1 August 2015].

[50] *Digital Archive of Letters in Flanders* (DALF) Available from: from: http://ctb.kantl.be/project/dalf/ [Accessed 1 August 2015]. The DALF project's guidelines for the annotation of correspondence is available from: http://ctb.kantl.be/project/dalf/dalfdoc/index.html.

eventually in ten printed volumes. Other letter projects which adopt TEI include *Vincent van Gogh: The Letters*[51] and *The Dolley Madison Digital Edition*.[52] Details of digital letter projects that are currently in progress can be found on the wiki site for the TEI Special Interest Group: Correspondence.[53] Indeed, as will be discussed in chapter five, since 2008 the TEI correspondence SIG have been developing a correspondence module that was recently included in the *TEI Guidelines*. However (and this returns to the opening argument of this review), all of the projects that have fed into the development of this correspondence module are primarily concerned with the annotation of letters by famous historical figures: linguists, authors, composers, essayists and politicians. There is a notable lack of projects that focus on letters by the popular classes, which, as will be discussed throughout this thesis, come with their own unique challenges.

While, generally speaking, the DALF and WeGA projects focus on the communicative function and structure of letters the *Corpus of Early English Correspondence* and the *Parsed Corpus of Early English Correspondence* (PCEEC),[54] produced by the University of Helsinki, is more concerned with the social variables of the letter writers themselves. The CEEC/PCEEC comprises 188 letter collections (12,000 letters) dating from c. 1403-1800. Many of these

[51] Jansen, L., Luijten, H. and Bakker, N. (2014) *Vincent van Gogh: The Letters*. Available from: http://www.vangoghletters.org/vg/ [Accessed 1 August 2015].
[52] Shulman, C. (2009) *The Dolley Madison Digital Edition*. Available from: http://rotunda.upress.virginia.edu/dmde/ [Accessed 1 August 2015].
[53] The TEI SIG: Correspondence: http://wiki.tei-c.org/index.php/SIG:Correspondence.
[54] *The Corpus of Early English Correspondence* (CEEC) (1993-present) was compiled by Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A. and Palander-Collin, M. in the Research Unit for Variation, Contacts and Change in English (VARIENG) at the University of Helsinki. More information about the corpus can be found on the project website: http://www.helsinki.fi/varieng/domains/CEEC.html [Accessed 1 May 2015]. The part-of-speech tagging of the *Parsed Corpus of Early English Correspondence* was carried out by Arja Nurmi (University of Helsinki) and the syntactic annotation by Ann Taylor (University of York). The sociolinguistic information for each correspondent was provided by the Helsinki team, and by Ann Taylor, assisted by Joanne Close, at York (see: http://www.ling.upenn.edu/histcorpora/annotation/index.html).

letters, particularly those dating from the early fifteenth to late seventeenth century, were written by 'members of the gentry, nobility or clergy', rather than the popular classes (Elspaβ 2012, p. 162).[55] Nevertheless, the resource serves as an excellent example of how sociobiographic and extra-linguistic information might be captured in a systematised way; the sender database, for instance, contains up to 27 parameters, including: year of birth, year of death, sex, rank, social mobility, education, and religion (Raumolin-Brunberg and Nevalainen 2007, p. 162). However, the annotation scheme used on the PCEEC project, although very sophisticated, does not conform to TEI standards. As such, it does not help to address the wider issue of developing a system of markup which can be used across all letter collections – including emigrant correspondence – allowing resources to interconnect; unless, of course, projects agree to use the system developed for the PCEEC, which is not ideal since the markup scheme was developed with a specific dataset and research questions in mind.

As an increasing number of digital emigrant letter projects are now surfacing, this, I would argue, is a good time for projects and disciplines to collaborate in order to develop a system of markup that will allow the various digital editions to interconnect. The *Swedish Emigrant Institute* in Vaxjo, for example, has 'an emigrants database' containing letters and diaries of Swedish migrants to America; however, at present, there does not appear to be online access to this resource.[56] Similarly, the sources (including letters) collected by the

[55] It is worth noting here that while sociobiographic and extra-linguistic information is often readily available for 'gentry, nobility or clergy', this is not the case for lower rank members of society. Whilst we can conjecture about who a particular person was (their age, whether they married, their occupations and so on) the lack of physical evidence relating to ordinary individuals can cause problems when it comes to capturing this type of information within the markup. See Nevalainen and Raumolin-Brunberg (1996, pp. 43-45) for more on this issue with reference to the CEEC.

[56] *Swedish Emigrant Institute*. Available from: http://www.kulturparkensmaland.se/1.0.1.0/14/2/

Centro di documentazione sulla storia dell'emigrazione trentina project, relating to individual stories of Italian migration, are not, as yet, freely available.[57] The *Wales-Ohio Project: Digitising the Archive of the Welsh in Ohio*, at the National Library of Wales, contains several letter series by Welsh migrants to Ohio in the nineteenth century. While digital transcriptions of the letters are not always available, images of the original manuscripts are provided and basic metadata has been captured, allowing at least some level of searchability – searches based on, for example, the language of the letter, the date (year), and the sender's location (county).[58] The *Glanidad* project, also at the National Library of Wales, provides a similar resource focusing on Welsh emigrants who settled in Patagonia in the late nineteenth century.[59] The *Dutch Immigrant Letters* project[60] at Calvin College has a collection of Dutch and German emigrant correspondence, which are referenced alphabetically on the website. Some of the references include PDF images of the manuscripts together with typed transcriptions of those manuscripts. The website also provides a detailed overview of what resources exist in the archive. Finally, The *Digitizing Immigrant Letters* project at the Immigration History Research Center at the University of Minnesota 'aims to make available on-line digitized letters from the IHRC Archives and other collections (private individuals, partner institutions) that were written between 1850 and 1970 both by immigrants (the so-called "America letters") and to

[Accessed 12 August 2015].

[57] *Centro di documentazione sulla storia dell'emigrazione trentina*. Available from: http://fondazione.museostorico.it/index.php/it/Progetti/Principali-ambiti-tematici-di-ricerca/Centro-di-documentazione-sulla-storia-dell-emigrazione-trentina [Accessed 12 August 2015].

[58] *Wales-Ohio Project* at the National Library of Wales. Available from: http://ohio.llgc.org.uk/index.php [Accessed 12 August 2015].

[59] *Glanidad* at the National Library of Wales. Available from: http://www.glaniad.com/index.php?lang=en [Accessed 12 August 2015].

[60] *Dutch Immigrant Letters* at Heritage Hall, Calvin College. Available from: http://www.calvin.edu/hh/letters/letters_main.htm [Accessed 12 August 2015].

immigrants (“homeland letters”)’.[61] The collection contains around 80 letters and includes images of the original manuscripts as well as digital transcriptions and translations of the letters. More extensive markup has been used to capture bibliographic metadata as well as metadata relating to the letters’ provenance and the participants involved in the act of communication. This metadata has been structured in a formalised way making it possible to convert the database contents into a TEI compliant format (as will be discussed in chapter five), thus allowing the collection to potentially interconnect with other resources.

In terms of resources which focus on Irish emigrant correspondence, the *Documenting Ireland: Parliament, People and Migration* (DIPPAM) project is perhaps one of the largest. This online archive of sources, hosted by Queen’s University, Belfast, consists of three principal databases: (a) Enhanced British Parliamentary Papers on Ireland (EPPI); (b) Irish Emigration Database (IED) and (c) Voices of Migration and Return (VMR). The IED component of the archive contains several thousand letters between Irish emigrants and their families and friends dating from 1700 to 1950. The texts themselves contain very little markup and although it is possible to carry out basic searches (all letters within a particular timespan or by a particular author, for instance) it is not possible to search the collection based on sociobiographic information (the author’s occupations or social status, for instance). Additionally, it is not possible to analyse the search outputs to notice patterns or trends within the letter content itself.

The *Corpus of Irish English Correspondence* (CORIECOR), at the University of Bergen, is a collection of emigration writings incorporating some of

[61] *Digitizing Immigrant Letters* at the Immigration History Research Centre, University of Minnesota. Available from: https://www.lib.umn.edu/ihrca/dil [Accessed 12 August 2015].

the letter data from the previously mentioned IED (largely late seventeenth to early twentieth century data) as well as a nineteenth century Irish-Argentinian collection (Amador-Moreno and McCafferty (2012)). The corpus compilers state that ‘eventually, it will also include migration correspondence from sources housed in archives and libraries in Ireland and abroad so that each twenty-year subperiod of the corpus [will] contain 200,000 words’ (Amador-Moreno et. al. 2015 *forth.*). At present these letters are not publicly available and there is no indication as to whether or not they will be annotated in accordance with TEI conventions, or indeed whether they will be annotated for sociobiographic information at all given that the purpose of this corpus is to investigate purely linguistic phenomena. However, already we can see issues of duplication: many of the CORIECOR letters are taken from the IED collection and both projects (CORIECOR and the IED) have different ways of organising and annotating their data. Additionally, while the IED data, as it stands, is clearly not suitable for linguistic analysis (as the letters are not in a text-searchable format), the CORIECOR data, unless it is annotated for sociobiographic information, will be of limited use to historians or sociolinguists. Furthermore, a significant number of the letters held in Kerby Miller’s archive (mentioned earlier in this review) also come from the IED. This situation (three projects potentially working with the same letters), I believe, presents a good case for developing a common system of markup for emigrant correspondence which would allow the three projects to interconnect and duplications to be easily identified.

To conclude, I would like to argue that the digital humanities, specifically the areas of corpus linguistics and text markup, offers one possible way in which emigrant letter collections might interconnect allowing for more nuanced and sophisticated corpus searches within and across editions. However, whilst an increasing number of letter projects are now utilising tools and techniques from the digital humanities to create fully marked-up and text searchable digital editions, many emigrant letter projects still appear to be working in isolation of one another, leading to the duplication of both effort and data. In this thesis I will explore how examining the emigrant letter within the broader framework of the digital humanities (specifically looking at corpus and computational methods of analysis and markup practices) might offer more collaborative and interdisciplinary opportunities for understanding histories from below.

CHAPTER ONE

The Lough Letters

Throughout this thesis I will be referring to the Lough family letters (letters home to Ireland, from the daughters who emigrated to America). While chapters two, three and four use corpus and computational methods to examine the content of these letters, chapters five and six focus on how to model metadata relating to emigrant correspondence, using letters from the Lough collection to demonstrate the proposed markup templates. The thesis focuses on just one letter series, using these to develop a set of solutions concerning methods of analysis and markup that can be applied across collections.

The Lough (pronounced Locke)[62] family letters are from Professor Kerby Miller's collection of Irish emigrant correspondence, held at the University of Missouri.[63] Significantly, these letters are drawn from a much larger body of Irish emigrant correspondence collected by Miller. Miller himself has explored this wider corpus in several pioneering works on Irish emigration (see, for instance, Miller (1985) and Miller et. al. (2003)) and his archive of over 5,000 letters has also been referred to by many scholars including Emmons (1990), Koos (2001), Bruce (2006), Corrigan (1992) and Noonan (2011). But the Lough family correspondence, which is a small but significant part of Miller's collection, has attracted less attention.

[62] Among the Irish relatives the spelling later became 'Locke' – very close to the Irish pronunciation of the name – and was written Lowe on some official documents.
[63] Professor Kerby Miller, Curators' Professor, Department of History, University of Missouri: http://history.missouri.edu/people/miller.html.

In the early 1950s, a few of the Lough letters were initially donated by Canice and Eilish O'Mahony of Dundalk, County Louth, to Arnold Schrier, then a graduate student at Northwestern University, now Professor Emeritus at the University of Cincinnati, who subsequently employed them, alongside other epistolary documents, in his 1958 book *Ireland and the Irish Emigration, 1850-1900*. In 1977-78 the rest of the Lough letters were donated to Miller by the O'Mahonys and by Edward Dunne and Kate Tynan of Portlaoise, County Laois. Both Miller and Schrier, who thereafter collaborated in researching Irish migration to America, made photocopies and transcriptions of these letters, and Miller returned the original manuscripts to their donors. It is this new material that Miller himself has offered the most detailed analysis of to date, in his 2008 study *Ireland and Irish America: Culture, Class, and Transatlantic Migration*, where he uses the Lough letters as part of a wider argument that 'Irish emigration was based on *family* – not individual – decisions: [on] choices by Irish parents as to which of their children to send or allow to go abroad first; and choices by Irish Americans as to which of their siblings, cousins, or other relatives to encourage and assist to emigrate and join them' (2008, p. 307).[64]

Indeed, this familial dynamic is clearly evident in the migration story of the Lough sisters. The post-famine period (circa. 1850s-1920s) was a time that saw a significant increase in female migration. Economic changes in Ireland, including declining wage earning capabilities due to the deindustrialisation of the Irish countryside, as well as changes in inheritance practices from partible to inpartible inheritance systems, leading, in turn, to changes in marriage trends

[64] The Lough sisters are also mentioned in Nolan (1989).

with women marrying ‘less frequently and at later ages than in the pre-famine past’ (Miller 1985, p. 3), contributed to ‘a massive post-famine emigration by young, unmarried women’ (ibid.). Between 1852 and 1921 the median age for female Irish emigrants was 21.2 and after 1880, young women, such as the Lough sisters, constituted the majority of the departing Irish (Miller 1985, p. 392). A small glimpse into the lives of these young women – their preoccupations, experiences, perceptions, and beliefs – can be found in the letters they wrote home to their families in Ireland.

The five Lough sisters - Elizabeth, Alice, Annie, Julia and Mary - came from a Roman Catholic family in Meelick, in what was then called Queen's County (now County Laois), Ireland.[65] The five sisters were daughters of Elizabeth McDonald Lough and James Lough who lived on a small holding consisting of two fields, one of which, according to family legend, was sold to pay for the sisters' passages. The Lough family were, according to Miller, not of the lowest class as both parents and daughters were able to write. Apart from Mary – the youngest – all the Lough sisters emigrated to America between 1870 and 1884. The sisters who emigrated were, in Miller's words, four ‘very dutiful, hard-working, and pious Irish female immigrants, who came to America at a time when Irish women comprised a majority of the Irish immigration to the U.S’;[66] the sisters remained very close both geographically and emotionally throughout their lives (the letters indicate that the sisters in America kept in touch via letters and the occasional visit to one another's homes).

[65] I am indebted to personal communications with Kerby Miller for the information that follows. See too Miller (2008, p. 316).

[66] This quote is taken from correspondence in the Lough file between Miller and Mrs Edward McKenna (one of the donors).

Elizabeth (sometimes referred to as Liz or Lizzie) Lough emigrated in 1870 to Winsted, Litchfield County, Connecticut, where she worked mainly as a seamstress. She married Dan Walsh,[67] who worked on a passenger train, and had five children (Tom, Alice, John William, Catherine Elizabeth and James). Elizabeth died in 1923,[68] but mention of her in her sisters' letters disappears after around 1912. Elizabeth was apparently the first sister to emigrate. She originally went to live with her aunt and uncle (from her mother's side) – George and Anne Burke – who preceded Elizabeth to America and may have paid for her passage tickets. It is likely that Elizabeth saved money to help bring out her other sisters.

Alice Lough (sometimes referred to as Alisha or Alicia) emigrated in the 1870s. Alice appears to have married before she emigrated – Miller has a copy of her marriage certificate dated 27 May 1875. In America, her husband, Edward Elliott, was an employee in a shop or factory that made coffins. Alice and her husband lived in Winsted, between 1870 and 1880, before then moving to Hampden County, Massachusetts in 1881, with several of their eventual seven children (Mary Elizabeth (born 14 August 1876), Edward, James (a railroad conductor, who died on 8 April 1918), William (who served in World War I), John, Alice and Phillip). Alice died on 23 September 1922.

Annie (sometimes referred to as Nan) Lough was the third sister to emigrate, in 1878; she lived in Winsted all her life, where she appears to have worked as a servant for a while. Annie married John McMahon on 9 June 1886 –

[67] According to the website *Findagrave*, Dan Walsh was possibly born in 1849 and died on 20 November 1896 (buried at St Joseph's, Winsted). Available from: http://www.findagrave.com [Accessed 1 August 2015].

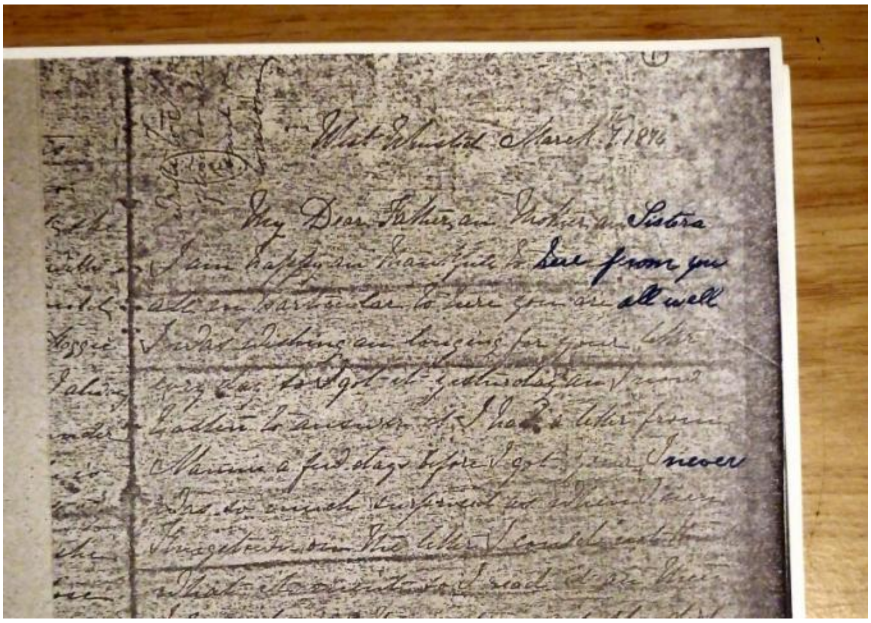
[68] According to *Findagrave*, Elizabeth ('Lizzie' Lowe Walsh) was born in November 1859 and died on 28 July 1923 (Lizzie, like her husband, was buried at St Joseph's, Winsted). Available from: http://www.findagrave.com [Accessed 1 August 2015].

a labourer or factory worker – however, she bore no children. Annie died in Winsted in 1935; her husband died on 18 September 1936.

And finally, Julia Lough came over in September 1884 – at the age of thirteen. After arriving, Julia lived with her sister Elizabeth and her brother-in-law Dan Walsh in Winsted between 1884 and 1894. In approximately 1895 she moved to Litchfield County, Connecticut, where she remained until at least 1927, the point when her letters stop. Julia was somewhat of a success story, working as a seamstress to begin with, then from the age of nineteen as an apprentice dressmaker, before becoming a professional dressmaker and opening up her own shop on Main Street, where she employed several members of staff. On 21 June 1897, at the age of twenty-five, Julia married a well-respected, Irish-born railroad engineer, Thomas McCarthy, with whom she had six children (although only one, Elise, is named in her letters). Julia died in Torrington, Litchfield County, Connecticut, on 22 February 1959; her husband died shortly after on 8 April 1959.

Mary Lough remained in Ireland with her mother and father. She married John Fitzpatrick and had four daughters. Besides Mary Lough, there may have been another Lough sister who stayed in Ireland, whose married name was Hickey. However, this is all that is mentioned in the Lough file.

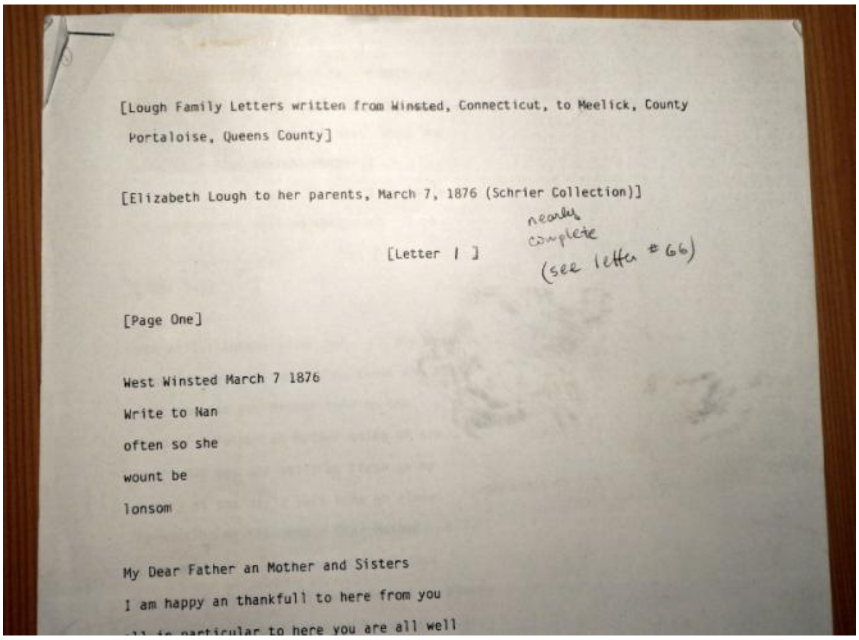
As mentioned previously, within the Lough collection, in most cases, there is a photocopy of the original manuscript together with one or more typed transcriptions of that manuscript. As an example, the first letter in the Lough series (a letter by Elizabeth Lough, dated 7 March 1876) comes in two parts. The first part (Part A) was donated to Schrier in the 1950s; the second part (Part B) was donated to Miller in the 1970s. (Miller was able to match these two fragments by comparing page sizes and ink colors.) Within this file there is a photocopy of Part A (see Figure 1.1) together with two typed transcriptions: one by Schrier (Figure 1.2) and one by Miller (Figure 1.3). Essentially, these transcriptions look the same; however, there are small (and sometimes significant) differences in Schrier's and Miller's transcriptions, which potentially affect meaning and interpretation. Table 1.1, for example, shows an extract from page four of Elizabeth's letter. Highlighted in yellow are differences between Schrier's transcription and Miller's transcription. While some of the differences are relatively minor (although, arguably, still important) – differences to do with spacing and capitalisation – there is, in Miller's transcription, a full line which is missing from Schrier's transcription.



West Winsted March 7 1876

My Dear Father an Mother an Sisters
I am happy an thankful to hear from you
all an particular to here you are all well
[illegible] wishing an longing for your letter
every day to get it yesterday an I [illegible]
[illegible] to answer it I had a letter from
Mamma a few days before I got yours I never
was so much surprised as when I [illegible]
[illegible] the letter I could [illegible]
what [illegible] I read it [illegible]

Figure 1.1: Photocopy of original manuscript (*ELC*, 7 March 1876)



[Lough Family Letters written from Winsted, Connecticut, to Meelick, County Portaloise, Queens County]

[Elizabeth Lough to her parents, March 7, 1876 (Schrier Collection)]

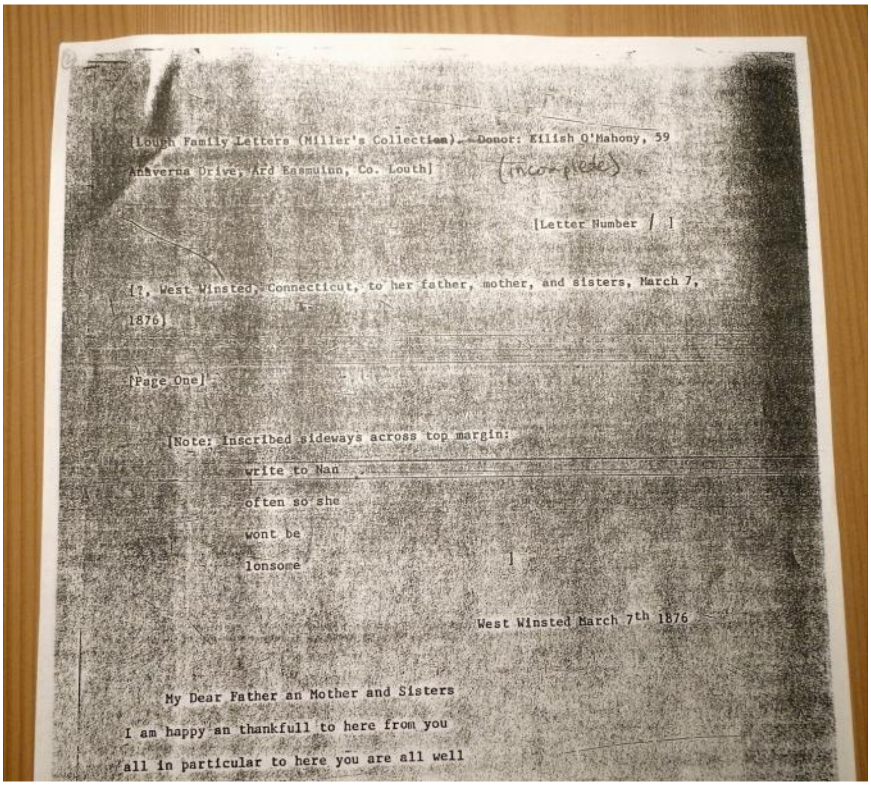
[Letter 1] nearly complete (see letter #66)

[Page One]

West Winsted March 7 1876
Write to Nan
often so she
wount be
lonsom

My Dear Father an Mother and Sisters
I am happy an thankfull to here from you
all in particular to here you are all well

Figure 1.2: Schrier’s transcription of Part A (*ELC*, 7 March 1876)



[Lough Family Letters (Miller's Collect~~ion)~~. Donor: Eilish O'Mahony, 59 Annaverna Drive, Ard Easmuinn, Co. Louth] (incomplete)

[Letter Number 1]

[?, West Winsted, Connecticut, to her father, mother, and sisters, March 7, 1876]

[Page One]

[Note: Inscribed sideways across top margin:
write to Nan
often so she
wont be
lonsome]

West Winsted March 7th 1876

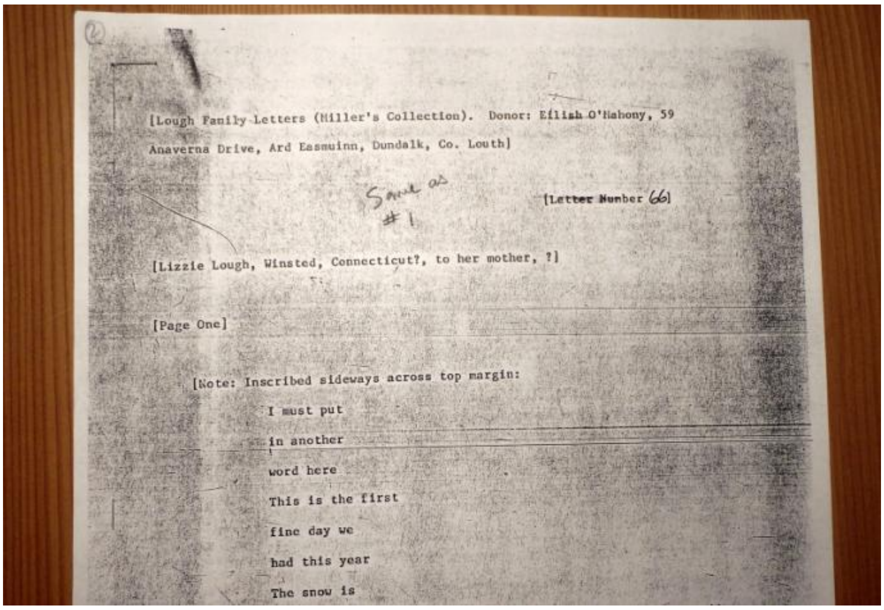
My Dear Father an Mother and Sisters
I am happy an thankfull to here from you
all in particular to here you are all well

Figure 1.3: Miller’s transcription of Part A (*ELC*, 7 March 1876)

Schrier's transcription	**Miller's transcription**
what is the matter with Maggie is she sickly or why is she no good I always thought she would be smart I wounder if she wount let me no what she is about Mary is she good I suppose she is most as big as me now who dose she look like I remember she had a little round nose did it grow long like mine yet let me know which of them can sew the best	what is the matter with Maggie is she sickly or why is she no good I always thought she would be smart I wounder if she Wount let me no what she is best at and you did not say anything about Mary is she good I suppose she is most as big as me now Who dose she look like I remember she had a little round nose did it grow long like mine yet let me no which of them can sew the best

Table 1.1: Differences in transcription (extract from *ELC*, 7 March 1876)

Additionally, within the file, there is Miller's transcription of Part B (note that a photocopy of the original manuscript is not available for the second part of the letter) – see Figure 1.4 – as well as an MS Word version of Part A and Part B combined (i.e. the complete letter), produced by Miller's research assistant (Figure 1.5). There are small differences between Miller's typed transcriptions and the RA's transcription. These could simply be typing errors, but without knowing for certain the reason for these differences, it is difficult to determine which version of Elizabeth's letter is most accurate. In chapter five, I will briefly look at issues to do with transcription practices and how this sort of information might be documented and modelled within the markup, thereby helping to produce more reliable data.



2

[Lough Family Letters (Miller's Collection). Donor: Eilish O'Mahony, 59 Anaverna Drive, Ard Easmuinn, Dundalk, Co. Louth]

Same as #1

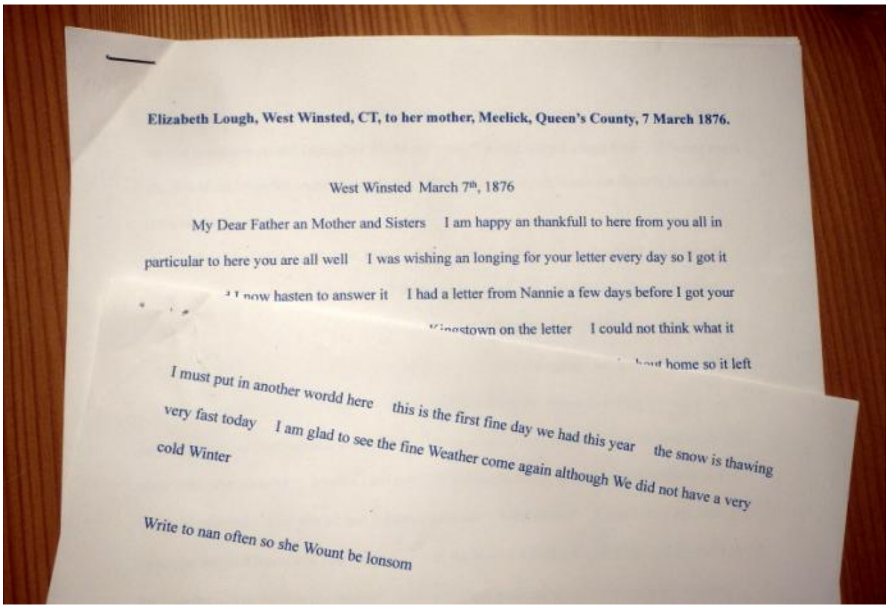
[Letter Number 66]

[Lizzie Lough, Winsted, Connecticut?, to her mother, ?]

[Page One]

[Note: Inscribed sideways across top margin:
I must put
in another
word here
This is the first
fine day we
had this year
The snow is

Figure 1.4: Miller's transcription of Part B (*ELC*, 7 March 1876)



Elizabeth Lough, West Winsted, CT, to her mother, Meelick, Queen's County, 7 March 1876.

West Winsted March 7th, 1876

My Dear Father an Mother and Sisters   I am happy an thankfull to here from you all in particular to here you are all well   I was wishing an longing for your letter every day so I got it
d I now hasten to answer it   I had a letter from Nannie a few days before I got your
Kingstown on the letter   I could not think what it
home so it left

I must put in another wordd here   this is the first fine day we had this year   the snow is thawing very fast today   I am glad to see the fine Weather come again although We did not have a very cold Winter

Write to nan often so she Wount be lonsom

Figure 1.5: Miller's RA's transcription of Parts A and B (*ELC*, 7 March 1876)

To create digital transcriptions of the Lough letters I have used Miller's transcriptions (rather than Miller's RA's). In cases where there are transcriptions by both Miller and Schrier – and where there are discrepancies between those transcriptions – I have cross-checked my digital transcription against a photocopy of the original manuscript, to try to iron out any inconsistencies. However, in cases where a photocopy of the original manuscript is not available, for the purpose of consistency, I have used Miller's interpretation.

For this thesis, I decided to manually transcribe the letters. However, for larger transcription projects OCR software might be a more appropriate option.[69] All that is required is a computer with good processing power and access to plenty of storage space so that images can be uploaded and outputs can be downloaded. The outputs can be produced in PDF or MS Word format, or, if the budget allows for more sophisticated scanning software, HTML or XML format.[70]

Potentially, using OCR software would save a lot of time, especially when working with large letter collections. However, there are problems. Although the scanning itself can be carried out relatively quickly, requiring little supervision, the outputs need to be manually checked, interpreted and corrected.

Bleeding text (where the ink spreads into the paper making the characters less distinct and identifiable) and thin paper (causing text on the reverse side to show through) can affect the quality of the scan. Variations in character and word spacing can also confuse the software as it tries to split words and understand each part separately, or combine words and understand the whole. Additionally, contrast between the text and the background can cause problems, as can complex and varying fonts as well as artefacts on the paper which can be

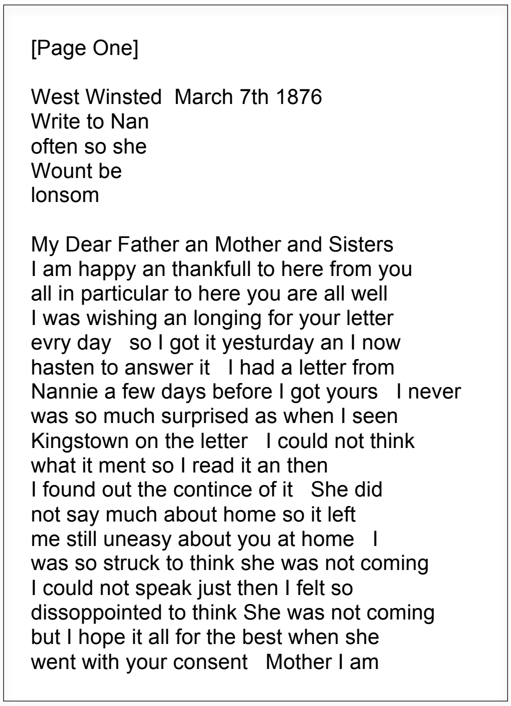
[69] The following three paragraphs, discussing the benefits and drawbacks of using OCR software, are based on email communications in December 2012 with Chris Mumby, Head of Commercial Delivery, from the National Archives, regarding the 'BT Connections' project (see: http://www.digitalarchives.bt.com/web/arena/research/hilary-emma). For this project I was a member of the Academic Team, responsible for digitising the correspondence component of the archive.
[70] This comparison table outlines some of the different OCR software that is available: http://en.wikipedia.org/wiki/Comparison_of_optical_character_recognition_software.

interpreted, incorrectly, as characters. Some software can be trained to recognise specific peculiarities within a letter series, thereby generating a bespoke database of repeated errors and, more importantly, the appropriate corrections that can be inserted in place of those errors. However these parameters would need to be reset every time a new letter series is scanned.

Finally, there is the issue of resolution. The resolution of the scan will depend on the quality of the original document, including the size of the font and the clarity of the letters. In short, the software needs to be able to recognise the characters – if the characters are too small (less than, say, font size 10), the original document will need to be enlarged. For poor quality texts (with issues such as those outlined above), a resolution of less than 300dpi will not be good enough to get acceptable results. Equally, however, a resolution that is too high can also cause problems: very high resolutions will pick up variations in the background that the software will try and interpret as text.

To summarise, projects will need to decide whether OCR scanning is right for them, based on factors such as the data they are working with (the quantity and quality) as well as the time and budget available to them. For my own research purposes, the time that would have been saved was not worth the effort, or financial cost, of sourcing and purchasing appropriate OCR software. Additionally, there is something to be said for manually transcribing the letters, as through this process the researcher gets a better sense of the data, which in turn can inform their markup decisions later on, as will be discussed in chapters five and six.

I created two digital versions of each letter in the Lough collection: one in MS Word, maintaining the layout of Miller's transcription, including any annotations, such as ‘[Note: Inscribed sideways in top left corner]’ and ‘[Page X]’, typically distinguished from the rest of the text by square brackets (see Figure 1.6); and one in Plain Text format (a suitable format for most corpus and content analysis software) where all formatting and annotations were removed leaving just the content of the letter (see Figure 1.7). Note that in Miller’s transcriptions he maintains word/character spacing and line breaks in accordance with the original manuscript. There are very few punctuation marks in the Lough letters; however, often extra space is given, which can be interpreted as indicating the start of a new sentence; Miller mirrors this in his transcriptions. Additionally, spelling variations have been maintained (correct or alternative spellings are not provided and the letters have not been ‘standardised’); in other words, Miller’s transcriptions (and my MS Word transcriptions) represent, as closely as possible, the language, structure and layout of the original manuscripts.

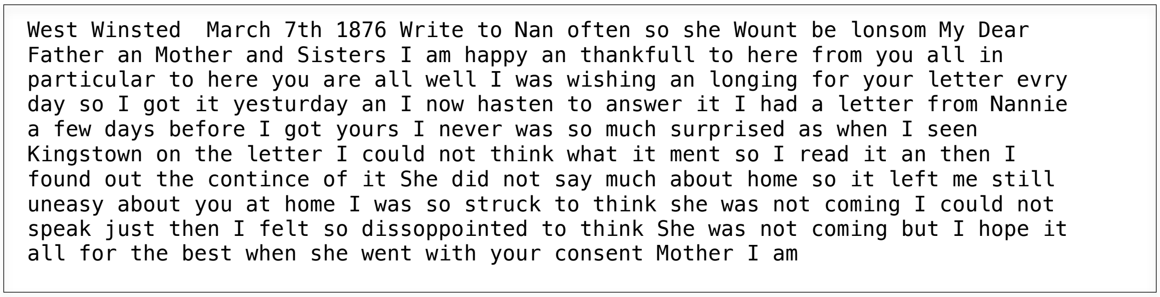


[Page One]

West Winsted March 7th 1876
Write to Nan
often so she
Wount be
lonsom

My Dear Father an Mother and Sisters
I am happy an thankfull to here from you
all in particular to here you are all well
I was wishing an longing for your letter
evry day   so I got it yesturday an I now
hasten to answer it   I had a letter from
Nannie a few days before I got yours   I never
was so much surprised as when I seen
Kingstown on the letter   I could not think
what it ment so I read it an then
I found out the contince of it   She did
not say much about home so it left
me still uneasy about you at home   I
was so struck to think she was not coming
I could not speak just then I felt so
dissoppointed to think She was not coming
but I hope it all for the best when she
went with your consent   Mother I am

Figure 1.6: Extract from the MS Word version of the Elizabeth Lough letter (*ELC*, 7 March 1876)



```
West Winsted  March 7th 1876 Write to Nan often so she Wount be lonsom My Dear
Father an Mother and Sisters I am happy an thankfull to here from you all in
particular to here you are all well I was wishing an longing for your letter evry
day so I got it yesturday an I now hasten to answer it I had a letter from Nannie
a few days before I got yours I never was so much surprised as when I seen
Kingstown on the letter I could not think what it ment so I read it an then I
found out the contince of it She did not say much about home so it left me still
uneasy about you at home I was so struck to think she was not coming I could not
speak just then I felt so dissoppointed to think She was not coming but I hope it
all for the best when she went with your consent Mother I am
```

Figure 1.7: Extract from the Plain Text version of the Elizabeth Lough letter (*ELC*, March 1876)

There are 100 documents in the Lough collection, dating from 7 March 1876 to 18 October 1928. Table 1.2[71] provides an overview of the data. As discussed in the previous section, documents LOUGH001 and LOUGH065 (see the first cell in the 'Ref' column) are, it is believed, extracts from the same letter by Elizabeth Lough. Where information is not known, fields have been marked 'Unknown', or '0', in the case of date columns, which require a numerical value.

'From: first' and 'From: surname' give the author's first name and surname. The 'To: first' and 'To: surname' columns provide the name of the recipient and the 'To: relation' column details the relationship between author/recipient – specifically whether the recipient/s is/are the author's parents (i.e. both mother and father), mother, sister, niece, nephew, or friend. The columns 'Town/City', 'State/County' and 'Country', in Table 1.2, provide address information for the author and/or recipient, and the date of the letter is captured as day/month/year; however, sometimes only partial or approximate dates are known.

Table 1.2 shows that most of the letters were sent from America (Connecticut – where Elizabeth, Julia and Annie were based, and Massachussetts

[71] Table 1.2 can also be accessed online, via *Google Docs*, by following this link: https://drive.google.com/file/d/0B9jGe2sY5rz8bWtrWUxDdFcwSEk/view?usp=sharing.

– where Alice was based) to Ireland (Meelick – where younger sister Mary and her parents lived). As is so often the case, the letters in the Lough collection are unidirectional and there is very limited information about the family members living in Ireland. Just three of the letters in the Lough collection were sent from Queenstown (now Cobh) on the County Cork coast of Ireland. Significantly, Annie sent one of these letters just as she was about to set sail for America in around 1878 and another was sent by Julia on 27 September 1884 when she was about to embark on the same journey. These letters very movingly capture some of the mixed feelings the sisters were experiencing as they prepared to leave Ireland for a new life in America.

There are four letters written by Elizabeth in the Lough collection (as previously noted one of these comes in two parts). The first letter is dated 7 March 1876 and the last letter is dated 31 January 1877. For Alice, there are nine letters; the first is dated 18 December 1889 and the last is dated 28 December 1914 – it should, however, be noted that five of the nine letters are not dated, so Alice's correspondence could span a much wider timeframe. Annie appears to write most frequently of all the Lough sisters, with 39 letters sent between 3 March 1890 and 18 October 1928. For 11 of Annie's letters the date is either unknown or ambiguous. Finally, there are 35 letters written by Julia. The first letter is dated 27 September 1884, from Queenstown, and the last is dated 17 March, in what Miller believes to be, 1919 or 1920. For several of Julia's letters the specific year is not known; however, from reading the content of the correspondence it has been possible to place them in an approximate timeframe.

While Table 1.2 provides information about what is in the Lough collection (the number of letters together with author, recipient and date information), Table 1.3[72] and Table 1.4[73] provide an overview of the content of the letters.

The online tool *Javascript Kit*[74] was used to count the number of words and characters (Table 1.3, columns B and C) in each letter. From this information I was able to calculate the average word length within each letter (column C). The average length of a letter in the Lough collection is 416.8 words, and the average word length is 4.8 characters. While *Javascript Kit* provides an overview of the number of words/characters within each letter, the text analysis software *LIWC* (Linguistic Inquiry and Word Count),[75] developed by researchers working in the fields of social psychology, language and health, provides a useful overview of what those words are. Column E, in Table 1.3, shows what percentage of the words, in each letter, are personal pronouns while columns F to J give a more detailed breakdown. The pronoun *I* is significantly more frequent than the other pronouns (*we*, *you*, *s/he*, and *they*), which one might expect in first person ego-documents such as personal letters.

Columns K, L and M give the percentage of past tense (*went*, *ran*, *had*), present tense (*is*, *does*, *hear*) or future tense (*will*, *going to*) verbs. For the most part it would appear that the Lough sisters wrote mainly in the present tense, an average of 10.97% across all letters, and – to a lesser extent – in the past tense (3.75%). Interestingly, the use of the future tense scores relatively low – 2.15%.

[72] Table 1.3 can be accessed online, via *Google Docs*, by following this link: https://drive.google.com/file/d/0B9jGe2sY5rz8dVloeDhaaGd1UUE/view?usp=sharing.
[73] Table 1.4 can be accessed online, via *Google Docs*, by following this link: https://drive.google.com/file/d/0B9jGe2sY5rz8ckl2R3NpNFljLUE/view?usp=sharing.
[74] *Javascript Kit*. Available from: http://www.javascriptkit.com/script/script2/charcount.shtml [Accessed 1 June 2011].
[75] Pennebaker, J. W., Booth, R. J. and Francis, M. E. (2007) *Linguistic Inquiry and Word Count* (LIWC2007). Available from: http://www.liwc.net [Accessed 1 June 2011].

As will be discussed in chapter four, a possible reason for this might be to do with themes such as homesickness and separation being particularly prevalent in emigrant letters, with the emigrant frequently looking back to past, shared events as a mechanism for reinforcing familial relationships in the present.

Columns N to R give percentages for social processes (words relating to family, friends and humans – *husband*, *friend*, *baby* etc.), affective processes (words expressing positive or negative emotion, as well as anxiety, anger and sadness – *nice*, *hurt*, *worried*, *annoyed*, *sad* etc.), cognitive processes (words expressing insight, causation, discrepancy, tentativeness, certainty, inhibition, inclusion and exclusion – *think*, *effect*, *should*, *maybe*, *always*, *stop*, *with*, *without*), perceptual processes (words which describe seeing, hearing and feeling – *see*, *listen*, *feel* etc.), biological processes (words relating to the body, health, sex and ingestion – *hands*, *flu*, *love*, *eat* etc.) and, finally, relativity (words relating to motion, time and space – *arrive*, *in*, *season* etc.). The higher percentages for cognitive processes (an average of 18.26%), social processes (15.98%) and relativity (12.6%) are quite striking when compared with affective processes (8.26%), perceptual processes (2%) and biological processes (1.21%), and may gesture to possible themes in the discourse relating to family (family relationships/dynamics) and the movement of people over time – although, of course, a more detailed analysis would need to be carried out.[76]

Columns T to AB give percentages for the number of words, within each letter, that relate to a particular topic, namely: space (with an average of 3.64%), time (7.08%), work (0.92%), achievement (0.95%), leisure (0.44%), home

[76] For a full breakdown of the categories, together with examples, see: http://www.liwc.net/descriptiontable1.php. Note that I have not looked at all of the linguistic categories within LIWC, specifically some of the grammatical categories, as my aim was to try to get an overall sense of the content of the letters, including possible topics and themes.

(0.85%), money (0.49%), religion (0.74%) and death (0.29%). Space and time appear to be key themes within the letters and, again, may be worth further investigation. And, finally, columns AC and AD give the percentage of positive and negative emotion words within each letter. Table 1.3 shows that the Lough letters contain significantly more positive emotion words (an average of 7.37%) compared with negative emotion words (an average of just 0.98%).

In Table 1.4 I have used the online corpus analysis and comparison tool Wmatrix[77] to identify the five most frequent semantic domains within each letter. The 'SEMTAG' column provides the semantic tag that has been assigned to a word or group of words, while the 'Semantic Field' column details what the SEMTAG represents (for example, SEMTAG 'T1.1.3' represents the semantic field 'Time: Future' (words such as *will*, *soon*, *going to*, *next day* etc.) and SEMTAG 'F1' represents the semantic field 'Food' (words such as *eat*, *meals*, *bake*, *supper*)). The 'Freq.' column gives the relative frequency of a particular SEMTAG.

In line with the LIWC results, 'Pronouns' is the most common semantic field within the Lough correspondence. The semantic field 'Existing' is also very frequent within the letters (words such as *is*, *am*, *are*, *was*, *were*, *being*) as is 'Personal Names'. Additionally, looking at columns I, L and O, 'Degree: Boosters' (words like *very*, *so*, *much*, *more*), 'Entire: Maximum' (words like all *any*, *each*, *every*), 'Getting and Possession' (words like *have*, *had*, *got*, *received*), 'Kin' (words like *mother*, *aunt*, *sister*, *father*), and 'Long, tall and wide' (words such as *long*) are the most frequently occurring semantic fields. A closer look at some of these words in context reveals that they are often part of the formulaic

[77] Rayson, P. (2009) *Wmatrix*. Lancaster University. Available from: http://ucrel.lancs.ac.uk/wmatrix/ [Accessed 1 June 2011].

and structural expressions typical of letter writing at the time – expressions such, *Dear* ***Mother*** *I* ***received*** *your* ***long*** *and welcome letter*, *from your* ***very*** *Affectionate* ***Sister***, and *We are* ***all*** *well here at present*. In other words, these frequently occurring semantic fields seem to be indicative of the letter writing genre.

There is much more that could be said about the findings shown in Tables 1.2, 1.3 and 1.4; however, the purpose of this chapter was simply to get a general overview of the Lough family correspondence. While it is clearly important to capture very basic information about the author/recipient, their locations and the date of the letters, having an initial sense of the content of the letters can also prove to be very useful. Using computational tools such as LIWC and Wmatix it was possible to identify trends and patterns in the language that might be worth further investigation and in the following chapters I will explore some of these initial observations in more detail.

Ref	From: first	From: surname	Town/City	State/County	Country	To: first	To: surname	To: relation	Town/City	State/County	Country	Day	Month	Year
LOUGH001 + LOUGH065	Elizabeth	Lough	Winsted	Connecticut	America	Elizabeth and James	Lough	Parents, Siblings	Meelick	Queens County	Ireland	7	March	1876
LOUGH002	Elizabeth	Lough	Winsted	Connecticut	America	Elizabeth and James	Lough	Parents	Meelick	Queens County	Ireland	21	August	1876
LOUGH003	Elizabeth	Lough	Winsted	Connecticut	America	Elizabeth and James	Lough	Parents, Siblings	Meelick	Queens County	Ireland	13	October	1876
LOUGH004	Elizabeth	Lough	Winsted	Connecticut	America	Elizabeth and James	Lough	Parents	Meelick	Queens County	Ireland	31	January	1877
LOUGH005	Julia	Lough	Queenstown	County Cork	Ireland	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	27	September	1884
LOUGH006	Julia	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	20	December	1884
LOUGH007	Alice	Lough	Westfield	Massachusetts	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	27	February	1888
LOUGH008	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	December	1888
LOUGH009	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	3	November	1889
LOUGH010	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	2	December	1889
LOUGH011	Alice	Lough	Westfield	Massachusetts	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	18	December	1889
LOUGH012	Annie	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother, Sister	Meelick	Queens County	Ireland	3	March	1890
LOUGH013	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	9	March	1890
LOUGH014	Margaret	Hourigan	Nenagh	Tipperary	Ireland	Mary	Lough Fitzpatrick	Cousin	Meelick	Queens County	Ireland	3	August	1890
LOUGH015	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	10	August	1890
LOUGH016	Unknown	Unknown	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother, Sister	Meelick	Queens County	Ireland	7	October	1890
LOUGH017	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	December	1890
LOUGH018	Julia	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	18	January	1891
LOUGH019	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	25	January	1891
LOUGH020	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	18	October	1891
LOUGH021	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	14	December	1891
LOUGH022	Annie	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	15	December	1891
LOUGH023	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	1	September	1892
LOUGH024	Annie	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	30	March	1893
LOUGH025	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	July	1893
LOUGH026	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	December	1893
LOUGH027	Julia	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	4	June	1894
LOUGH028	Julia	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	November	1894
LOUGH029	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	10	October	1893
LOUGH030	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	17	March	1895
LOUGH031	Julia	Lough	Torrington	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	August	1895
LOUGH033	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	18	May	1899
LOUGH034	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	30	March	1891
LOUGH036	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	16	February	1901
LOUGH037	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	21	September	1901
LOUGH038	Unknown	Unknown	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	8	December	1901
LOUGH039	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	10	December	1902
LOUGH040	Alice	Lough	Westfield	Massachusetts	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	12	August	1904
LOUGH041	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	3	April	1906
LOUGH042	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	20	June	1906
LOUGH043	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	30	November	1906
LOUGH044	Alice	Lough	Westfield	Massachusetts	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	7	January	1910
LOUGH045	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	12	December	1912
LOUGH046	Unknown	Unknown	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	29	January	1913
LOUGH047	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	8	December	1913
LOUGH048	Annie	Lough	Winsted	Connecticut	America	Unknown	Unknown	Niece	Meelick	Queens County	Ireland	11	December	1914
LOUGH049	Alice	Lough	Westfield	Massachusetts	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	28	December	1914
LOUGH050	Ian	Pigott	Maidenhead	Berkshire	England	Mary	Lough Fitzpatrick	Friend (ML)	Meelick	Queens County	Ireland	2	June	1917
LOUGH051	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	31	April	1918
LOUGH052	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	6	May	1918

Table 1.2: Lough Corpus - Overview

Ref	From: first	From: surname	Town/City	State/County	Country	To: first	To: surname	To: relation	Town/City	State/County	Country	Day	Month	Year
LOUGH053	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	14	July	1918
LOUGH054	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	14	August	1919
LOUGH055	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	1	December	1919
LOUGH056	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	29	September	1925
LOUGH057	Julia	Lough	Torrington	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	9	November	1927
LOUGH058	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	28	March	1928
LOUGH059	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	18	October	1928
LOUGH060	Annie	Lough	Queenstown	County Cork	Ireland	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	18	June	0
LOUGH061	Annie	Lough	Winsted	Connecticut	America	James	Unknown	Nephew	Unknown	Unknown	Unknown	4	November	1910
LOUGH062	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	1884
LOUGH063	Unknown	Unknown	Unknown	Unknown	Unknown	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH064	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	1892-1893
LOUGH066	Annie	Lough	Unknown	Unknown	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH067	Annie	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH068	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	11	May	pre-1892
LOUGH069	Annie	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	29	October	1891
LOUGH070	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	May	1893
LOUGH071	Unknown	Unknown	Unknown	Unknown	Unknown	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH072	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	1889-1894
LOUGH073	Unknown	Unknown	Unknown	Unknown	Unknown	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH074	Julia	Lough	Torrington	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	1889-1894
LOUGH075	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	3	September	1893
LOUGH076	Julia	Lough	Torrington	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	25	March	1894
LOUGH077	Alice	Lough	Westfield	Massachusetts	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH079	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	1884-1894
LOUGH080	Mag	Hourigan?	Unknown	Unknown	Unknown	Mary	Lough Fitzpatrick	Cousin?	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH081	Annie	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	23	March	1892
LOUGH082	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	0	Unknown	0
LOUGH083	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH084	Unknown	Unknown	Unknown	Unknown	Unknown	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH085	Julia	Lough	Torrington	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	24	May	1893-1894
LOUGH086	Julia	Lough	Queenstown	County Cork	Ireland	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	8	July	1895
LOUGH087	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH088	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH089	Julia	Lough	Winsted	Connecticut	America	Elizabeth	McDonald Lough	Mother	Meelick	Queens County	Ireland	0	Unknown	1889-1890
LOUGH090	Alice	Lough	Westfield	Massachusetts	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	8	March	0
LOUGH091	Annie	Lough	Unknown	Unknown	Unknown	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH092	Annie	Lough	Unknown	Unknown	Unknown	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH093	Alice	Lough	Westfield	Massachusetts	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	10	December	0
LOUGH094	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH095	Alice	Lough	Westfield	Massachusetts	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	30	March	0
LOUGH096	Alice	Lough	Westfield	Massachusetts	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	16	October	0
LOUGH098	Annie	Lough	Unknown	Unknown	Unknown	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	0	Unknown	0
LOUGH100	Annie	Lough	Winsted	Connecticut	America	Katie	Unknown	Niece	Meelick	Queens County	Ireland	21	March	1920
LOUGH101	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	31	March	1924
LOUGH102	Julia	Lough	Torrington	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	17	March	1919-1920
LOUGH103	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	7	December	1919 or 1929
LOUGH104	Annie	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	21	March	1920
LOUGH105	Julia	Lough	Winsted	Connecticut	America	Mary	Lough Fitzpatrick	Sister	Meelick	Queens County	Ireland	21	March	1893

Table 1.2: Lough Corpus - Overview (cont.)

Ref	Words	Characters	Av. Word Length	Personal Pronouns %	I %	We %	You %	S/he %	They %	Past %	Present %	Future %	Social Processes %	Affective Processes %	Cognitive Processes %	Perceptual Processes %	Biological Processes %	Relativity %	Space %	Time %	Work %	Achieve-ment %	Leisure %	Home %	Money %	Religion %	Death %	Positive Emotion %	Negative Emotion %
LOUGH001 + LOUGH065	1443	6553	4.54	19.94	10.70	0.28	2.78	5.42	0.76	4.24	10.70	2.22	15.15	5.00	15.64	2.08	1.11	10.98	2.92	6.46	0.63	1.25	0.56	1.11	0.42	0.35	0.21	4.10	0.97
LOUGH002	571	2712	4.75	15.62	6.60	1.04	3.99	3.65	0.35	5.73	7.99	1.91	16.84	5.03	12.50	1.74	1.39	10.76	2.08	8.33	1.39	1.39	0.35	0.00	0.52	0.69	0.00	5.21	0.00
LOUGH003	297	1401	4.72	16.11	7.38	1.68	2.35	3.36	1.34	2.68	10.40	2.35	15.44	6.71	13.76	1.34	0.67	11.07	1.34	7.72	0.34	1.01	0.00	0.34	0.00	0.34	0.00	5.37	1.34
LOUGH004	539	2508	4.65	15.77	8.91	0.74	3.15	2.78	0.19	5.01	7.24	1.48	13.73	5.01	15.03	1.11	1.67	11.13	2.97	6.86	1.30	1.48	0.19	1.11	0.74	0.19	0.19	3.53	1.48
LOUGH005	40	218	5.45	19.05	11.90	0.00	7.14	0.00	0.00	4.76	9.52	7.14	11.90	9.52	19.05	2.38	0.00	16.67	0.00	16.67	0.00	0.00	0.00	0.00	0.00	0.00	2.38	9.52	0.00
LOUGH006	519	2472	4.76	17.66	6.53	1.15	4.22	4.99	0.77	3.84	10.17	1.15	19.77	6.91	16.89	1.34	0.77	15.93	3.26	9.98	0.19	0.58	0.19	0.58	0.77	0.58	0.00	6.33	0.58
LOUGH007	376	1833	4.88	19.54	11.05	0.26	4.63	0.77	2.83	1.80	12.34	2.31	16.97	5.40	22.11	1.03	0.77	9.77	2.83	5.40	0.00	0.26	0.00	1.03	0.00	0.51	0.00	4.88	0.77
LOUGH008	342	1636	4.78	18.08	8.16	0.87	9.04	0.00	0.00	1.46	11.95	1.75	16.62	6.12	19.83	2.33	0.00	12.54	2.04	7.87	0.87	0.58	0.29	0.58	1.46	0.29	0.00	5.25	0.87
LOUGH009	444	2174	4.90	11.43	2.47	2.24	2.02	2.02	2.69	4.26	8.74	0.00	15.70	5.16	13.90	3.14	1.12	10.99	3.36	4.93	1.57	1.12	0.22	0.22	0.45	2.24	1.12	4.93	0.22
LOUGH010	436	2234	5.12	15.98	7.13	1.08	4.75	2.38	0.65	1.94	12.53	2.16	15.98	9.29	16.20	0.86	1.08	12.74	2.81	6.70	0.65	0.43	0.43	0.22	1.73	1.51	0.00	8.64	0.65
LOUGH011	525	2560	4.88	16.38	6.21	1.51	5.08	1.32	2.26	2.07	10.92	2.45	15.82	5.08	20.34	1.51	1.13	13.75	5.08	6.03	1.13	0.94	0.19	1.13	0.75	0.94	0.56	4.14	0.94
LOUGH012	469	2248	4.79	14.26	4.68	1.70	6.60	1.28	0.00	4.47	8.94	1.91	16.17	8.51	18.94	2.77	1.28	13.40	3.83	8.09	0.21	0.85	0.21	0.85	0.00	1.28	0.43	7.45	1.06
LOUGH013	463	2269	4.90	17.85	7.10	1.29	4.30	4.52	0.65	3.87	11.40	2.37	18.71	8.17	16.56	1.72	1.51	12.47	3.01	7.74	1.72	1.08	0.65	0.22	0.22	0.43	0.00	7.53	0.86
LOUGH014	297	1406	4.73	17.11	10.74	0.34	4.03	1.01	1.01	2.35	8.72	3.02	12.75	8.72	16.78	1.01	1.34	13.09	3.69	7.05	0.00	0.00	0.34	1.01	0.00	1.34	0.00	8.05	0.67
LOUGH015	366	1769	4.83	15.26	5.45	0.27	4.36	4.63	0.54	4.90	13.62	1.63	17.44	8.99	16.62	3.00	1.36	11.17	3.00	6.27	1.09	1.09	0.27	1.36	0.00	0.27	0.00	7.63	1.36
LOUGH016	482	2299	4.77	14.43	4.95	0.21	4.95	2.89	1.44	4.33	10.52	2.68	14.43	8.87	16.49	2.68	0.62	9.69	3.92	4.74	1.24	1.03	0.21	1.24	0.21	0.41	0.41	7.22	1.86
LOUGH017	350	1692	4.83	19.09	7.98	1.14	7.69	1.42	0.85	1.14	12.25	3.13	19.09	11.11	19.94	1.99	1.14	11.11	1.99	7.12	0.57	0.57	0.00	0.28	0.85	1.99	0.00	10.54	0.57
LOUGH018	348	1653	4.75	15.47	9.74	0.57	3.44	1.43	0.29	4.30	8.31	2.01	12.89	5.44	18.34	1.72	0.29	12.89	2.58	7.45	3.15	1.72	0.29	0.57	2.01	0.29	0.00	5.16	0.29
LOUGH019	351	1637	4.66	14.53	7.12	1.99	4.27	1.14	0.00	5.98	12.82	0.85	10.26	7.98	19.09	1.42	1.14	14.81	4.56	8.55	0.00	0.85	0.57	0.00	0.57	1.71	1.14	5.41	2.56
LOUGH020	317	1475	4.65	14.47	6.29	0.00	4.09	3.77	0.31	1.57	16.98	2.83	14.47	15.09	12.89	1.26	1.89	10.38	2.20	6.60	0.63	0.31	0.00	0.63	0.00	0.31	0.00	13.84	1.26
LOUGH021	300	1489	4.96	15.79	6.25	0.33	2.96	4.61	1.64	2.63	11.18	2.96	16.78	9.54	20.07	2.63	0.66	10.53	3.62	6.25	0.99	1.32	0.00	2.30	0.00	0.99	0.00	8.55	0.99
LOUGH022	452	2106	4.66	18.06	7.05	0.88	4.41	4.41	1.32	3.08	12.33	2.42	16.08	9.03	18.28	1.32	1.10	12.78	3.74	7.27	1.32	1.32	0.00	0.88	0.88	0.88	0.00	7.93	1.10
LOUGH023	396	1938	4.89	17.63	8.06	0.25	2.02	6.30	1.01	3.02	12.85	1.01	17.88	8.56	16.12	2.27	0.76	14.86	3.78	8.56	2.27	0.25	0.76	1.01	0.25	0.00	0.00	7.05	1.51
LOUGH024	945	4489	4.75	14.57	5.39	0.95	4.01	2.53	1.69	3.48	10.03	2.11	14.89	6.23	17.42	2.22	1.69	14.68	3.70	8.45	0.42	0.21	0.11	0.84	0.21	1.48	0.63	5.49	0.74
LOUGH025	340	1616	4.75	14.96	4.69	1.17	2.35	5.57	1.17	4.11	15.54	1.76	17.89	8.80	19.06	4.11	1.17	12.02	3.23	7.04	1.76	2.05	0.00	0.88	0.00	0.59	0.00	7.92	0.88
LOUGH026	451	2194	4.86	15.04	5.53	0.66	5.09	2.65	1.11	3.10	10.40	1.55	17.70	11.28	17.48	1.77	1.33	10.62	3.76	4.87	0.22	0.66	0.88	1.33	0.88	1.33	0.00	9.73	1.55
LOUGH027	736	3526	4.79	14.50	7.59	0.81	4.07	0.95	1.08	4.34	11.52	1.76	12.33	6.37	16.40	3.25	1.22	14.91	5.42	7.59	1.36	1.22	0.41	0.95	0.95	0.68	0.00	5.83	0.54
LOUGH028	469	2310	4.93	15.32	5.74	1.28	4.26	2.55	1.49	3.19	12.13	2.13	16.17	7.66	16.81	1.49	0.85	12.98	3.83	8.09	0.85	0.85	0.43	0.21	1.28	1.49	0.64	6.60	1.49
LOUGH029	334	1568	4.69	19.05	11.01	0.30	4.76	2.68	0.30	2.08	11.61	1.49	14.29	8.93	17.86	2.38	0.60	12.50	2.98	8.04	0.30	1.19	0.89	1.19	0.30	0.30	0.00	8.93	0.30
LOUGH030	579	2775	4.79	16.35	4.65	1.89	3.96	4.30	1.55	4.30	10.84	1.89	18.24	10.33	20.48	2.41	1.38	10.33	2.58	7.06	0.52	1.03	0.34	0.52	0.00	0.69	0.17	9.29	1.03
LOUGH031	416	1964	4.72	18.18	9.33	0.24	5.74	2.39	0.48	6.94	8.61	2.63	13.16	8.61	22.49	2.39	0.72	8.85	3.11	3.83	1.20	0.48	0.00	0.00	0.96	0.24	0.00	7.66	0.96
LOUGH033	510	2395	4.70	16.73	7.78	0.58	5.64	2.14	0.58	5.45	8.95	2.33	17.90	9.73	21.98	1.36	1.75	11.87	2.92	6.81	0.97	1.36	0.39	0.78	0.39	0.00	1.17	8.37	1.56
LOUGH034	225	1051	4.67	17.70	7.96	0.44	5.31	3.98	0.00	2.21	12.83	3.10	17.26	9.73	21.24	1.33	0.88	11.50	2.21	6.19	0.00	0.00	0.44	0.44	0.44	0.44	0.00	9.29	0.44
LOUGH036	386	1864	4.83	15.72	6.70	1.03	5.41	0.77	1.80	4.90	9.79	2.58	17.01	7.47	18.81	1.29	0.77	9.28	2.84	5.41	2.06	1.03	0.52	0.77	0.00	0.00	0.00	6.44	1.03
LOUGH037	415	2040	4.92	14.05	5.48	1.19	4.76	2.38	0.24	4.05	10.24	1.67	18.10	7.86	15.95	1.90	1.19	10.71	3.57	5.24	1.43	0.95	0.71	0.71	0.95	0.00	0.71	6.90	0.95
LOUGH038	269	1289	4.79	15.87	5.54	2.95	3.69	2.95	0.74	6.64	8.86	1.48	17.71	9.23	21.40	1.48	0.74	9.23	2.21	3.69	0.37	0.37	0.74	1.48	0.00	0.00	0.00	8.49	0.74
LOUGH039	354	1725	4.87	15.97	5.88	1.40	4.48	1.12	3.08	1.96	12.61	2.24	17.09	9.52	21.29	0.84	1.40	12.61	3.08	6.72	2.80	0.84	0.56	1.12	0.28	2.24	0.28	9.24	0.28
LOUGH040	416	2049	4.93	18.72	8.77	0.95	3.08	2.37	3.55	4.50	8.77	1.90	17.54	6.64	15.17	2.13	0.71	13.27	5.21	7.11	1.42	1.42	1.18	1.18	0.71	1.42	0.47	6.40	0.47
LOUGH041	414	2011	4.86	14.90	6.01	0.96	4.33	1.68	1.92	0.96	12.26	2.64	15.62	9.13	18.51	1.20	1.92	15.62	3.85	10.10	2.88	1.44	0.72	0.72	0.48	0.48	0.00	8.89	0.24
LOUGH042	321	1536	4.79	13.58	5.25	1.23	4.01	0.62	2.47	3.70	10.19	4.01	18.21	7.72	16.67	2.78	0.31	12.35	2.47	7.41	0.93	0.31	1.23	0.62	0.31	0.00	0.31	7.10	0.93
LOUGH043	276	1343	4.87	17.27	7.19	0.00	6.47	1.44	2.16	2.52	10.43	3.24	21.94	8.63	14.75	2.52	0.72	11.15	2.16	5.40	0.00	0.72	0.72	0.72	0.00	0.72	0.00	8.99	0.00
LOUGH044	356	1735	4.87	18.01	7.48	0.83	4.16	3.32	2.22	4.16	9.70	1.39	19.94	8.31	12.47	1.66	0.83	16.34	4.16	10.25	0.55	0.55	0.55	1.11	0.55	0.00	0.00	8.31	0.55
LOUGH045	622	2934	4.72	13.48	3.85	0.32	4.98	4.01	0.32	2.89	12.52	2.89	15.09	8.51	18.46	2.57	1.44	9.15	2.41	4.33	1.93	2.09	0.48	0.48	0.32	0.48	0.16	6.58	1.93
LOUGH046	335	1586	4.73	13.95	6.23	0.59	3.26	3.56	0.30	2.67	13.95	2.67	12.76	5.64	16.02	4.15	0.89	14.54	3.86	10.68	0.59	1.19	0.00	1.48	0.59	0.89	0.00	5.34	0.30
LOUGH047	498	2353	4.72	12.05	5.02	1.20	3.41	2.01	0.40	2.41	10.84	1.81	12.45	9.64	20.08	2.01	0.60	14.46	3.61	9.44	0.40	1.41	0.40	1.81	0.80	0.60	0.00	9.04	0.80
LOUGH048	387	1872	4.84	13.08	6.41	1.03	2.82	1.79	1.03	3.85	11.54	2.05	13.33	8.21	24.10	1.03	1.54	14.87	4.62	8.97	0.51	0.51	0.26	0.77	0.77	0.77	0.51	6.15	2.31
LOUGH049	267	1276	4.78	18.08	6.27	0.74	5.17	5.54	0.37	7.38	9.23	1.85	20.30	7.38	14.02	4.43	1.11	11.44	3.69	7.38	0.00	0.74	0.74	0.74	0.00	0.37	0.37	5.54	1.85
LOUGH050	209	1073	5.13	13.81	4.76	0.48	4.76	3.33	0.48	6.19	7.14	1.90	12.86	12.38	15.24	1.43	1.43	14.29	5.24	7.62	0.00	0.48	0.00	0.95	0.00	0.00	0.48	8.10	4.29
LOUGH051	630	3050	4.84	17.35	5.84	0.79	7.26	2.52	0.95	5.99	10.25	1.58	18.61	8.68	19.56	1.42	1.10	12.46	3.63	6.94	0.63	0.79	0.32	1.58	0.32	0.47	0.16	7.26	1.42
LOUGH052	869	4160	4.79	12.83	3.58	1.73	3.70	3.01	0.81	3.93	9.94	1.62	15.38	8.32	19.08	1.39	1.16	12.02	4.51	5.66	0.81	1.73	0.69	1.04	1.50	0.69	1.16	6.71	1.85
LOUGH053	848	4113	4.85	14.64	4.25	0.94	5.90	1.89	1.65	4.25	9.09	1.89	17.47	8.15	19.48	2.01	0.94	12.75	4.49	7.20	0.59	1.42	0.83	1.30	0.12	0.94	0.94	6.14	2.13
LOUGH054	819	3906	4.77	16.77	6.49	1.35	4.65	2.82	1.47	5.02	9.18	1.22	15.54	8.57	20.56	1.35	0.86	13.46	3.92	6.73	1.35	0.61	0.98	1.96	0.49	0.24	0.61	7.22	1.35
LOUGH055	473	2263	4.78	15.68	5.72	0.85	6.99	1.69	0.42	3.18	10.38	1.91	16.31	9.32	19.92	1.06	0.64	11.65	5.72	4.24	1.27	1.69	0.64	0.85	0.21	0.42	0.00	9.32	0.42
LOUGH056	228	1088	4.77	13.36	5.17	1.29	4.31	1.72	0.86	3.88	9.91	2.59	15.95	8.19	18.97	1.29	1.29	12.07	3.02	6.47	1.29	0.43	0.43	2.59	0.00	0.00	0.00	8.62	0.00
LOUGH057	349	1773	5.08	18.75	5.11	1.99	6.25	5.40	0.00	5.97	8.52	1.99	21.31	9.94	16.48	0.57	2.27	9.09	3.12	3.98	0.28	0.85	0.00	0.28	0.57	4.83	1.70	7.39	2.56
LOUGH058	199	935	4.70	17.41	6.47	1.00	8.46	1.00	0.50	3.48	9.95	2.99	15.42	14.93	18.91	0.50	0.50	9.45	1.99	4.98	1.00	1.99	1.00	1.49	0.50	2.49	0.00	13.93	1.00
LOUGH059	845	3915	4.63	16.67	4.02	1.06	4.26	6.15	1.18	4.96	9.69	2.60	18.32	7.92	18.44	1.77	0.95	12.77	3.55	6.97	0.83	1.54	0.47	0.59	0.35	1.06	1.42	6.62	1.42
LOUGH060	337	1528	4.53	16.22	9.73	2.95	2.06	0.00	1.47	4.72	9.73	2.65	12.09	7.08	18.88	1.77	1.18	13.27	4.13	6.19	0.59	0.88	0.00	0.29	0.00	0.29	0.00	5.31	1.77

Table 1.3: LIWC Results

Ref	Words	Characters	Av. Word Length	Personal Pronouns %	I %	We %	You %	S/he %	They %	Past %	Present %	Future %	Social Processes %	Affective Processes %	Cognitive Processes %	Perceptual Processes %	Biological Processes %	Relativity %	Space %	Time %	Work %	Achieve-ment %	Leisure %	Home %	Money %	Religion %	Death %	Positive Emotion %	Negative Emotion %
LOUGH061	516	2578	5.00	14.42	3.90	0.19	4.09	5.46	0.78	6.04	7.99	1.75	18.71	8.58	16.37	2.73	1.36	12.48	5.85	4.48	1.56	0.78	0.97	1.36	0.00	0.97	0.97	7.99	0.78
LOUGH062	98	472	4.82	15.31	10.20	2.04	3.06	0.00	0.00	1.02	10.20	3.06	13.27	7.14	21.43	1.02	1.02	17.35	8.16	10.20	1.02	0.00	0.00	0.00	0.00	0.00	0.00	7.14	0.00
LOUGH063	215	1057	4.92	18.14	8.37	0.00	6.51	1.86	1.40	6.05	12.09	2.33	16.74	8.37	23.72	4.19	1.40	10.70	1.86	8.84	0.47	0.00	0.47	0.00	0.47	0.47	0.00	5.12	3.26
LOUGH064	321	1551	4.83	13.08	5.61	1.87	0.31	4.98	0.31	4.98	6.85	0.93	14.33	4.67	14.64	4.05	3.12	18.07	5.30	9.35	0.93	1.25	0.31	0.00	0.31	1.25	1.56	4.36	0.31
LOUGH066	204	980	4.80	17.16	3.92	0.49	7.35	4.90	0.49	3.43	9.80	2.45	23.53	12.25	20.10	0.49	0.49	11.76	1.96	8.82	0.49	0.98	0.49	0.98	0.49	2.45	0.00	12.25	0.00
LOUGH067	630	2952	4.69	13.81	5.40	0.79	2.38	2.54	2.70	4.76	9.37	2.22	16.35	6.03	18.41	1.75	1.43	11.11	5.24	4.44	0.16	0.63	0.48	0.79	0.32	0.16	0.32	4.29	1.75
LOUGH068	400	1895	4.74	18.25	6.50	1.25	4.50	5.50	0.50	5.00	12.25	2.25	16.75	7.00	16.00	3.25	0.50	15.00	3.50	8.25	1.50	1.25	0.00	1.25	0.25	0.50	0.00	6.50	0.50
LOUGH069	1032	4869	4.72	14.63	4.07	1.45	4.65	3.97	0.48	3.39	12.69	2.03	17.54	8.53	18.60	2.33	1.36	11.72	3.49	5.91	0.97	1.26	0.48	0.78	0.39	0.29	0.10	7.95	0.68
LOUGH070	305	1443	4.73	16.72	6.89	1.31	2.30	5.90	0.33	4.26	12.13	1.97	17.38	7.21	16.39	3.28	0.98	13.44	2.95	7.87	0.66	1.31	0.66	0.66	0.00	1.31	0.00	7.21	0.00
LOUGH071	193	931	4.82	15.54	7.25	0.00	1.55	6.74	0.00	1.55	12.95	3.63	17.10	9.33	16.58	2.59	1.55	15.03	4.66	7.25	1.55	1.55	0.52	1.04	0.00	1.04	0.00	8.81	0.52
LOUGH072	259	1264	4.88	16.99	6.95	0.00	3.47	5.41	1.16	4.63	10.04	1.54	18.15	3.86	18.92	1.93	0.77	15.83	4.25	8.49	1.54	1.93	0.00	1.16	0.39	0.00	0.39	2.70	1.16
LOUGH073	286	1297	4.53	14.69	2.80	3.85	3.85	3.85	0.35	8.04	7.34	2.80	17.13	4.20	26.57	1.75	0.00	12.94	5.59	5.24	0.70	1.05	0.00	1.05	0.35	2.10	0.70	2.45	1.75
LOUGH074	354	1724	4.87	14.93	6.48	1.69	4.51	0.85	1.41	4.23	10.70	0.56	14.93	9.58	17.46	3.10	0.28	10.14	3.66	6.48	1.69	2.25	0.56	0.56	0.56	1.13	0.00	9.30	0.28
LOUGH075	356	1708	4.80	12.89	6.44	0.84	3.08	1.12	1.40	2.24	11.76	2.24	13.17	7.00	19.33	2.80	1.12	12.61	3.92	6.44	1.68	1.12	0.56	1.40	1.68	0.84	0.00	6.72	0.28
LOUGH076	183	942	5.15	11.89	5.41	1.62	3.24	1.08	0.54	2.16	15.14	0.54	12.97	9.73	18.38	1.08	2.16	19.46	5.95	9.73	0.54	1.08	1.08	0.54	0.54	0.54	0.00	9.73	0.00
LOUGH077	106	542	5.11	15.09	4.72	2.83	7.55	0.00	0.00	0.94	5.66	3.77	20.75	8.49	12.26	0.00	1.89	18.87	7.55	4.72	0.00	0.00	0.00	1.89	0.00	0.94	0.00	8.49	0.00
LOUGH079	190	951	5.01	16.23	6.81	0.52	5.76	3.14	0.00	0.52	15.18	2.62	18.85	8.38	15.18	4.19	2.09	13.61	4.19	6.81	1.05	1.57	0.52	0.52	1.05	0.00	0.00	8.38	0.00
LOUGH080	418	1933	4.62	12.83	5.94	0.24	1.43	4.99	0.24	3.80	12.35	2.38	11.40	7.60	15.20	1.43	3.33	9.98	2.38	5.94	1.19	0.48	0.24	0.00	2.14	0.24	0.00	5.23	2.38
LOUGH081	975	4632	4.75	14.46	5.95	1.23	3.69	2.36	1.23	3.28	12.10	2.46	14.15	9.03	18.87	2.36	1.03	14.46	3.38	8.21	0.41	0.72	0.72	1.33	0.21	0.41	0.31	8.41	0.82
LOUGH082	23	104	4.52	13.04	13.04	0.00	0.00	0.00	0.00	0.00	21.74	0.00	0.00	4.35	17.39	4.35	0.00	13.04	4.35	4.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	4.35	0.00
LOUGH083	416	1934	4.65	13.70	5.77	0.48	2.64	2.88	1.92	4.81	8.65	2.16	15.87	7.45	19.71	1.68	2.16	12.74	3.12	6.97	0.48	0.96	0.72	1.68	0.00	0.24	0.24	6.97	0.72
LOUGH084	397	1878	4.73	16.62	7.30	0.25	2.27	3.53	3.27	3.27	11.34	3.02	13.60	8.06	26.70	2.02	1.51	12.59	3.02	6.55	1.01	1.76	0.50	0.76	0.00	0.25	0.25	6.80	1.26
LOUGH085	477	2383	5.00	14.88	7.23	0.62	3.93	1.65	1.45	2.89	13.84	1.24	13.02	10.95	14.88	1.86	1.65	11.57	3.93	5.79	1.24	0.62	0.41	0.62	0.62	0.83	0.00	10.33	0.83
LOUGH086	44	235	5.34	13.33	6.67	0.00	4.44	0.00	2.22	0.00	13.33	4.44	17.78	17.78	22.22	0.00	2.22	8.89	2.22	4.44	0.00	0.00	0.00	0.00	4.44	0.00	0.00	15.56	2.22
LOUGH087	240	1148	4.78	11.67	5.00	0.83	1.67	3.75	0.42	4.58	14.58	1.67	13.75	8.75	21.25	1.25	1.67	11.25	5.42	4.58	1.25	3.33	0.83	0.83	2.08	0.00	0.42	8.33	0.42
LOUGH088	379	1787	4.72	12.37	6.05	0.53	4.47	0.53	0.79	1.84	13.95	2.63	12.37	9.47	18.68	1.32	2.11	13.68	3.42	8.16	0.00	0.79	0.26	0.53	1.32	1.05	0.00	8.16	1.32
LOUGH089	487	2288	4.70	15.43	8.23	0.41	2.47	4.32	0.00	2.67	11.32	3.29	12.76	5.56	18.11	2.06	1.23	16.87	3.70	10.70	0.82	0.82	0.82	0.82	0.21	0.41	0.00	4.73	0.82
LOUGH090	262	1281	4.89	16.85	7.12	0.75	4.49	4.12	0.37	9.36	6.74	1.12	18.35	10.86	15.36	2.25	1.12	12.73	2.62	8.99	0.75	1.12	1.12	1.50	0.00	0.75	2.25	8.99	1.87
LOUGH091	479	2266	4.73	12.50	4.38	0.62	4.58	1.25	1.67	2.50	10.62	2.50	15.62	5.00	18.54	1.25	2.29	13.54	4.38	7.71	2.50	1.04	0.42	0.42	0.21	0.42	0.21	4.17	1.04
LOUGH092	457	2156	4.72	15.75	5.47	2.41	4.16	1.75	1.97	5.03	11.38	1.97	15.54	8.32	22.98	1.31	1.31	14.66	4.81	8.53	1.53	1.09	0.22	0.66	0.22	0.22	0.22	7.22	1.09
LOUGH093	440	2123	4.83	19.50	9.07	1.13	4.76	2.72	1.81	2.72	15.42	2.04	17.01	9.30	17.46	3.40	1.13	10.20	2.49	6.80	1.13	0.68	0.68	1.13	0.23	0.00	0.00	9.52	0.45
LOUGH094	189	904	4.78	11.28	2.05	3.08	5.13	0.51	0.51	2.05	7.18	1.03	17.44	7.18	23.08	2.05	2.05	14.36	4.62	8.72	0.00	1.03	0.00	0.00	0.51	1.54	0.00	7.18	0.00
LOUGH095	325	1542	4.74	17.82	9.06	0.60	3.32	1.21	3.63	3.93	13.60	0.91	14.50	5.74	16.62	1.51	1.21	11.18	3.63	5.14	1.81	1.51	0.60	1.51	0.00	0.91	0.00	5.44	0.30
LOUGH096	302	1435	4.75	19.47	4.95	1.98	5.28	6.27	0.99	6.27	7.92	2.31	22.77	7.59	15.18	1.32	1.65	10.89	2.64	6.27	1.65	0.66	0.66	1.32	0.33	0.00	0.99	7.59	0.99
LOUGH098	295	1398	4.74	16.61	4.07	1.69	4.75	4.07	2.03	1.69	14.92	1.36	17.97	7.80	24.07	2.03	1.02	11.19	3.73	6.10	1.02	0.68	0.34	1.02	0.00	3.05	0.00	6.78	1.02
LOUGH100	459	2160	4.71	17.57	6.51	0.00	7.38	3.69	0.00	4.12	11.28	1.95	19.09	9.76	21.48	2.82	0.87	9.54	2.60	6.07	0.43	1.08	0.22	0.87	0.00	0.22	0.22	8.89	1.74
LOUGH101	523	2499	4.78	14.26	3.61	1.52	4.94	3.04	1.14	4.37	10.84	1.71	15.59	10.08	20.15	1.71	1.52	12.36	5.89	4.94	0.38	0.38	0.57	1.33	0.38	0.57	0.57	7.41	2.66
LOUGH102	331	1729	5.22	15.36	5.12	1.51	6.02	0.60	2.11	4.82	9.34	1.51	17.17	6.93	14.46	2.11	3.31	15.66	3.61	9.34	0.30	0.30	1.81	1.20	1.20	1.81	0.00	6.63	0.30
LOUGH103	388	1873	4.83	13.14	5.41	1.03	4.38	1.03	1.29	4.90	9.54	2.32	11.86	7.99	17.78	1.80	0.77	16.24	6.96	6.19	1.29	0.77	0.77	1.03	0.52	1.80	0.26	7.73	0.26
LOUGH104	641	3140	4.90	11.18	3.42	1.09	2.95	2.33	1.40	3.73	7.92	1.86	12.89	7.45	17.70	2.80	1.09	12.89	4.35	7.76	1.09	0.47	0.78	0.47	0.00	0.16	0.62	6.21	1.40
LOUGH105	423	2016	4.77	17.65	8.94	0.47	6.82	1.18	0.24	2.82	14.59	2.59	16.94	9.88	19.76	3.29	0.71	12.00	1.18	9.41	0.47	1.18	0.24	0.24	0.00	0.24	0.00	10.12	0.00
Total	41268	197587	476.53	1547.17	641.42	100.20	430.22	272.01	103.29	371.48	1086.22	213.01	1582.14	817.51	1807.78	198.42	119.36	1254.70	359.94	700.70	90.63	93.94	43.81	83.69	48.48	73.29	28.93	729.54	97.17
Average	416.85	1996	4.8	15.63	6.48	1.01	4.35	2.75	1.04	3.75	10.97	2.15	15.98	8.26	18.26	2.00	1.21	12.67	3.64	7.08	0.92	0.95	0.44	0.85	0.49	0.74	0.29	7.37	0.98

Table 1.3: LIWC Results (cont.)

Ref	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.
LOUGH001 + LOUGH065	Z8	Pronouns	22.73	A3+	Existing	4.83	A9+	Getting and possession	2.74	Z6	Negative	2.45	S4	Kin	2.38
LOUGH002	Z8	Pronouns	20.83	A3+	Existing	4.71	S4	Kin	3.26	T1.3	Time: Period	2.72	Z1	Personal names	2.54
LOUGH003	Z8	Pronouns	20.77	A3+	Existing	4.58	Z1	Personal names	2.82	S4	Kin	2.82	A9+	Getting and possession	2.82
LOUGH004	Z8	Pronouns	19.15	A3+	Existing	4.26	Z1	Personal names	3.48	Q1.2	Substances and materials: Liquid	3.09	A9+	Getting and possession	2.51
LOUGH005	Z8	Pronouns	26.32	A7+	Likely	10.53	T1.3	Time: Period	5.26	S4	Kin	5.26	A3+	Existing	5.26
LOUGH006	Z8	Pronouns	19.96	Z1	Personal names	4.73	A3+	Existing	3.50	A9+	Getting and possession	2.26	Z6	Negative	2.06
LOUGH007	Z8	Pronouns	22.10	S4	Kin	2.96	A3+	Existing	2.96	Z1	Personal names	2.43	A9+	Getting and possession	2.16
LOUGH008	Z8	Pronouns	20.85	A3+	Existing	4.23	Q1.2	Paper documents and writing	3.93	M6	Location and direction	2.11	A13.3	Degree: Boosters	1.81
LOUGH009	Z8	Pronouns	13.26	A3+	Existing	6.28	A9+	Getting and possession	2.79	Z1	Personal names	2.56	B5	Clothes and personal belongings	2.33
LOUGH010	Z8	Pronouns	20.14	A9+	Getting and possession	3.24	Z1	Personal names	3.24	S9	Kin	3.01	A3+	Existing	3.01
LOUGH011	78	Pronouns	17.98	A3+	Existing	3.16	Z1	Personal names	1.98	N3.7+	Long, tall and wide	1.98	S4	Kin	1.78
LOUGH012	Z8	Pronouns	17.70	A3+	Existing	4.14	T1.3	Time: Period	3.68	Q1.2	Substances and materials: Liquid	3.45	A9+	Getting and possession	2.99
LOUGH013	Z8	Pronouns	20.63	A3+	Existing	4.71	A9+	Getting and possession	4.04	Z1	Personal names	2.91	Q1.2	Paper documents and writing	2.91
LOUGH014	Z8	Pronouns	20.56	T1.1.3	Time: Future	3.14	A3+	Existing	2.79	A5.1+	Evaluation: Good	2.79	M1	Moving, coming and going	2.79
LOUGH015	Z8	Pronouns	18.18	A3+	Existing	6.45	A13.3	Degree: Boosters	2.64	Q1.2	Substances and materials: Liquid	2.64	Z1	Personal names	2.35
LOUGH016	Z8	Pronouns	18.10	A3+	Existing	6.90	S4	Kin	3.02	A9+	Getting and possession	2.37	M6	Location and direction	2.16
LOUGH017	Z8	Pronouns	21.69	A3+	Existing	4.82	T1.1.3	Time: Future	3.31	S9	Kin	2.71	Q1.2	Paper documents and writing	2.71
LOUGH018	Z8	Pronouns	20.18	Q1.2	Paper documents and writing	3.98	A3+	Existing	3.67	A9+	Getting and possession	3.06	Z1	Personal names	2.75
LOUGH019	Z8	Pronouns	18.58	A3+	Existing	5.57	A13.3	Degree: Boosters	2.79	S9	Kin	2.17	A9+	Getting and possession	2.17
LOUGH020	Z8	Pronouns	18.60	A3+	Existing	7.64	A13.3	Degree: Boosters	3.65	A9+	Getting and possession	3.65	A5.1+	Evaluation: Good	3.32
LOUGH021	Z8	Pronouns	16.96	A3+	Existing	4.84	A9+	Getting and possession	3.46	T1.1.3	Time: Future	3.11	S4	Kin	2.42
LOUGH022	Z8	Pronouns	20.97	A3+	Existing	3.92	Z1	Personal names	3.00	A9+	Getting and possession	2.53	X2.6+	Expected	2.30
LOUGH023	Z8	Pronouns	19.03	A3+	Existing	6.43	Z1	Personal names	2.95	Q1.2	Substances and materials: Liquid	2.68	N5.1+	Entire; maximum	2.41
LOUGH024	Z8	Pronouns	17.59	A3+	Existing	4.45	N5.1+	Entire; maximum	2.12	A9+	Getting and possession	2.12	T1.3	Time: Period	2.00
LOUGH025	Z8	Pronouns	18.27	A3+	Existing	7.12	Z1	Personal names	4.33	A9+	Getting and possession	4.33	A13.3	Degree: Boosters	3.10
LOUGH026	Z8	Pronouns	17.18	A3+	Existing	6.35	Z1	Personal names	2.59	N3.7+	Long, tall and wide	2.59	S9	Religion and the supernatural	2.35
LOUGH027	Z8	Pronouns	16.72	A3+	Existing	5.38	A9+	Getting and possession	2.18	Z1	Personal names	2.03	T1.3	Time: Period	1.74
LOUGH028	Z8	Pronouns	17.55	A3+	Existing	4.85	T1.3	Time: Period	3.00	M6	Location and direction	3.00	Z1	Personal names	2.31
LOUGH029	Z8	Pronouns	22.67	A3+	Existing	4.97	N5.1+	Entire; maximum	2.48	A13.3	Degree: Boosters	1.86	S4	Kin	1.55
LOUGH030	Z8	Pronouns	18.77	A3+	Existing	7.22	X2.6+	Expected	2.53	Z1	Personal names	2.35	A13.3	Degree: Boosters	2.35
LOUGH031	Z8	Pronouns	22.42	A3+	Existing	5.29	A7+	Likely	4.79	A13.3	Degree: Boosters	2.27	B5	Clothes and personal belongings	2.02
LOUGH033	Z8	Pronouns	20.49	A3+	Existing	4.30	Q1.2	Paper documents and writing	3.07	Z1	Personal names	2.66	T1.1.3	Time: Future	2.46
LOUGH034	Z8	Pronouns	21.53	Z1	Personal names	4.31	A9+	Getting and possession	3.35	Q1.2	Substances and materials: Liquid	3.35	X2.6+	Expected	2.87
LOUGH036	Z8	Pronouns	17.69	A3+	Existing	4.83	Q1.2	Paper documents and writing	3.49	N3.7+	Long, tall and wide	3.22	M6	Location and direction	3.22
LOUGH037	Z8	Pronouns	16.42	A3+	Existing	6.72	Z1	Personal names	3.48	N3.7+	Long, tall and wide	3.48	S4	Kin	2.24
LOUGH038	Z8	Pronouns	19.31	A3+	Existing	4.63	Z1	Personal names	3.09	N3.7+	Long, tall and wide	3.09	M6	Location and direction	2.70
LOUGH039	Z8	Pronouns	17.40	A3+	Existing	4.42	N5.1+	Entire; maximum	3.83	S9	Kin	2.65	X2.6+	Expected	2.65
LOUGH040	Z8	Pronouns	19.85	A3+	Existing	4.77	S4	Kin	3.27	A13.3	Degree: Boosters	3.02	N5.1+	Entire; maximum	2.51
LOUGH041	Z8	Pronouns	16.07	A3+	Existing	4.59	T1.1.3	Time: Future	3.32	T1.3	Time: Period	2.81	Z1	Personal names	2.55
LOUGH042	Z8	Pronouns	19.11	Z1	Personal names	4.14	T1.1.3	Time: Future	3.50	A3+	Existing	3.18	A13.3	Degree: Boosters	2.87
LOUGH043	Z8	Pronouns	21.64	T1.1.3	Time: Future	4.10	A3+	Existing	3.73	Q1.2	Substances and materials: Liquid	3.73	M2	Putting, pulling, pushing, transporting	2.99
LOUGH044	Z8	Pronouns	20.65	A3+	Existing	5.90	S4	Kin	3.54	T1.1.2	Time: Present; simultaneous	2.95	A9+	Getting and possession	2.95
LOUGH045	Z8	Pronouns	17.40	A3+	Existing	5.74	A13.3	Degree: Boosters	3.04	Z1	Personal names	2.70	A9+	Getting and possession	2.53
LOUGH046	Z8	Pronouns	16.98	A3+	Existing	7.86	A13.3	Degree: Boosters	3.77	T1.3	Time: Period	2.52	T1.1.3	Time: Future	2.52
LOUGH047	Z8	Pronouns	16.52	A3+	Existing	4.13	Z1	Personal names	2.83	X2.6+	Expected	2.61	A13.3	Degree: Boosters	2.61
LOUGH048	Z8	Pronouns	26.98	A3+	Existing	3.27	S4	Kin	2.72	T1.1.3	Time: Future	2.72	N5.1+	Entire; maximum	2.45
LOUGH049	Z8	Pronouns	19.85	A3+	Existing	6.49	A13.3	Degree: Boosters	5.34	Q1.2	Substances and materials: Liquid	2.67	A9+	Getting and possession	2.29
LOUGH050	Z8	Pronouns	16.67	A3+	Existing	5.39	A13.3	Degree: Boosters	2.94	T1.3	Time: Period	2.45	A5.1+	Evaluation: Good	2.45

Table 1.4: Wmatrix Results

Ref	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.	SEM TAG	Semantic Field	Freq.
LOUGH052	Z8	Pronouns	15.22	A3+	Existing	3.62	N5.1+	Entire; maximum	1.93	Z1	Personal names	1.69	A13.3	Degree: Boosters	1.69
LOUGH053	Z8	Pronouns	17.68	A3+	Existing	4.20	N5.1+	Entire; maximum	2.60	A13.3	Degree: Boosters	2.47	A9+	Getting and possession	2.10
LOUGH054	Z8	Pronouns	18.94	A3+	Existing	3.89	Z1	Personal names	2.72	A7+	Likely	1.95	M1	Moving, coming and going	1.82
LOUGH055	Z8	Pronouns	17.14	A3+	Existing	4.18	N5.1+	Entire; maximum	2.86	A9+	Getting and possession	2.42	A7+	Likely	2.20
LOUGH056	Z8	Pronouns	13.24	N5.1+	Entire; maximum	4.11	A3+	Existing	4.11	Z1	Personal names	3.20	T1.1.3	Time: Future	3.20
LOUGH057	Z8	Pronouns	20.12	S9	Religion and the supernatural	4.50	A3+	Existing	3.60	T1.3	Time: Period	3.30	E2+	Like	2.70
LOUGH058	Z8	Pronouns	19.27	A3+	Existing	5.21	N5.1+	Entire; maximum	3.12	E2+	Like	2.60	Z1	Personal names	2.60
LOUGH059	Z8	Pronouns	19.11	A3+	Existing	4.71	Z1	Personal names	2.23	A13.3	Degree: Boosters	2.23	A7+	Likely	1.99
LOUGH060	Z8	Pronouns	17.51	A3+	Existing	3.26	A1.1.1	General actions / making	2.67	M1	Moving, coming and going	2.08	T1.1.3	Time: Future	2.08
LOUGH061	Z8	Pronouns	16.01	A3+	Existing	4.16	Z1	Personal names	2.70	M6	Location and direction	2.70	T1.3	Time: Period	2.08
LOUGH062	Z8	Pronouns	15.62	A3+	Existing	7.29	E2+	Like	4.17	T1.1.2	Time: Present; simultaneous	4.17	Z1	Personal names	4.17
LOUGH063	Z8	Pronouns	20.59	A3+	Existing	4.90	S4	Kin	3.43	N3.7+	Long, tall and wide	2.94	A1.1.2	Damaging and destroying	2.94
LOUGH064	Z8	Pronouns	15.89	A3+	Existing	3.31	Z1	Personal names	2.98	M1	Moving, coming and going	2.65	T1.3	Time: Period	1.99
LOUGH066	Z8	Pronouns	19.49	Q1.2	Paper documents and writing	4.10	Z1	Personal names	3.59	E2+	Like	3.08	S4	Kin	3.08
LOUGH067	Z8	Pronouns	16.95	A3+	Existing	5.03	S4	Kin	2.52	A13.3	Degree: Boosters	2.35	M6	Location and direction	2.01
LOUGH068	Z8	Pronouns	22.10	A3+	Existing	5.12	M6	Location and direction	2.70	T1.1.3	Time: Future	2.43	A7+	Likely	2.16
LOUGH069	Z8	Pronouns	17.98	A3+	Existing	5.62	A13.3	Degree: Boosters	2.35	Z1	Personal names	2.25	A5.1+	Evaluation: Good	1.94
LOUGH070	Z8	Pronouns	18.03	A3+	Existing	4.76	A9+	Getting and possession	3.74	T1.3	Time: Period	2.72	Z1	Personal names	2.38
LOUGH071	Z8	Pronouns	19.23	A3+	Existing	6.04	T1.1.3	Time: Future	3.85	A7+	Likely	3.30	Q2.1	Speech: Communicative	3.30
LOUGH072	Z8	Pronouns	19.59	A3+	Existing	3.67	S4	Kin	2.86	A13.3	Degree: Boosters	2.86	Z1	Personal names	2.45
LOUGH073	Z8	Pronouns	18.84	A3+	Existing	5.07	A7+	Likely	3.26	S9	Kin	2.54	S4	Kin	2.54
LOUGH074	Z8	Pronouns	17.43	A3+	Existing	4.28	Z1	Personal names	3.67	N3.7+	Long, tall and wide	3.06	A9+	Getting and possession	2.45
LOUGH075	Z8	Pronouns	14.59	A3+	Existing	5.17	N5.1+	Entire; maximum	2.74	Z6	Negative	2.74	N1	Numbers	2.43
LOUGH076	Z8	Pronouns	11.76	Z1	Personal names	4.12	A3+	Existing	4.12	T1.3	Time: Period	3.53	M1	Moving, coming and going	3.53
LOUGH077	Z8	Pronouns	16.33	Q1.2	Paper documents and writing	5.10	T1.1.3	Time: Future	4.08	E2+	Like	3.06	S4	Kin	3.06
LOUGH079	Z8	Pronouns	17.74	A3+	Existing	4.84	B5	Clothes and personal belongings	4.30	T1.1.3	Time: Future	3.23	X2.6+	Expected	2.69
LOUGH080	Z8	Pronouns	14.96	A3+	Existing	5.49	A9+	Getting and possession	2.74	Z1	Personal names	2.49	T1.1.3	Time: Future	2.24
LOUGH081	Z8	Pronouns	15.94	A3+	Existing	5.31	M6	Location and direction	2.71	N3.7+	Long, tall and wide	2.06	Z1	Personal names	2.06
LOUGH082	Z8	Pronouns	13.64	A3+	Existing	9.09	A7+	Likely	4.55	N3.7+	Long, tall and wide	4.55	X2.2+	Knowledgeable	4.55
LOUGH083	Z8	Pronouns	16.20	A3+	Existing	5.82	Z1	Personal names	4.30	S4	Kin	2.78	M6	Location and direction	2.78
LOUGH084	Z8	Pronouns	18.77	A3+	Existing	5.36	N5.1+	Entire; maximum	3.75	A13.3	Degree: Boosters	3.49	A7+	Likely	2.95
LOUGH085	Z8	Pronouns	17.29	A3+	Existing	5.99	F1	Food	2.44	A13.3	Degree: Boosters	2.22	A9+	Getting and possession	2.22
LOUGH086	Z8	Pronouns	16.67	Z1	Personal names	7.14	E2+	Like	7.14	N3.7+	Long, tall and wide	7.14	T1.1.3	Time: Future	4.76
LOUGH087	Z8	Pronouns	12.23	A3+	Existing	5.24	N5.1+	Entire; maximum	3.93	G1.1	Government	3.06	T1.1.3	Time: Future	2.62
LOUGH088	Z8	Pronouns	14.84	A3+	Existing	3.57	T1.1.3	Time: Future	2.47	Z1	Personal names	2.20	T1.3	Time: Period	2.47
LOUGH089	Z8	Pronouns	18.04	A3+	Existing	3.34	Q1.2	Paper documents and writing	2.45	M1	Moving, coming and going	2.45	Z1	Personal names	2.23
LOUGH090	Z8	Pronouns	18.85	A3+	Existing	5.00	A13.3	Degree: Boosters	2.69	T1.3	Time: Period	2.31	S4	Kin	2.31
LOUGH091	Z8	Pronouns	15.08	A3+	Existing	4.88	Z1	Personal names	3.77	N3.7+	Long, tall and wide	3.77	T1.1.3	Time: Future	2.00
LOUGH092	Z8	Pronouns	18.75	A3+	Existing	6.02	M6	Location and direction	3.24	A5.1+	Evaluation: Good	2.31	N5.1+	Entire; maximum	2.08
LOUGH093	Z8	Pronouns	21.88	A3+	Existing	5.41	N5.1+	Entire; maximum	3.29	X3.4	Sensory: Sight	2.35	M6	Location and direction	2.35
LOUGH094	Z8	Pronouns	13.98	M6	Location and direction	3.23	T1.3	Time: Period	3.23	A3+	Existing	3.23	E2+	Like	2.15
LOUGH095	Z8	Pronouns	20.00	A3+	Existing	5.08	A9+	Getting and possession	4.13	M6	Location and direction	3.17	A13.3	Degree: Boosters	2.54
LOUGH096	Z8	Pronouns	21.65	A3+	Existing	5.84	S4	Kin	3.44	E2+	Like	3.09	A13.3	Degree: Boosters	3.09
LOUGH098	Z8	Pronouns	18.51	A3+	Existing	4.63	S9	Religion and the supernatural	3.56	A9+	Getting and possession	3.56	N5.1+	Entire; maximum	3.56
LOUGH100	Z8	Pronouns	21.09	A3+	Existing	5.22	A13.3	Degree: Boosters	3.40	S4	Kin	3.40	E2+	Like	2.27
LOUGH101	Z8	Pronouns	16.70	A3+	Existing	5.43	Z1	Personal names	3.02	M6	Location and direction	2.82	A13.3	Degree: Boosters	2.82
LOUGH102	Z8	Pronouns	16.72	A9+	Getting and possession	2.89	A3+	Existing	2.89	N3.7+	Long, tall and wide	2.89	S9	Religion and the supernatural	2.57
LOUGH103	Z8	Pronouns	16.26	A3+	Existing	4.88	M6	Location and direction	2.17	T1.1.3	Time: Future	2.17	N5.1+	Entire; maximum	2.17
LOUGH104	Z8	Pronouns	14.64	A3+	Existing	2.80	M6	Location and direction	2.30	A9+	Getting and possession	2.14	Z1	Personal names	1.97
LOUGH105	Z8	Pronouns	20.15	A3+	Existing	5.90	A13.3	Degree: Boosters	3.19	A7+	Likely	2.70	A9+	Getting and possession	2.70

Table 1.4: Wmatrix Results (cont.)

CHAPTER TWO

Case study: Using corpus methods as a way into emigrant letter collections: exploring language and gender

Introduction

This first case study builds on the body of quantitative research outlined in the literature review. Looking specifically at language and gender, it proposes a complementary methodology which is based on the theories and techniques of corpus linguistics for examining emigrant letters – a methodology which attempts to bridge the gap between the content observed and the conclusions that are later drawn from that content; and one which moves between the quantitative and the qualitative and back again.[78] Whilst recognising that linguistic choices will reveal something about the context of situation, the context of culture and how the author construes events and perceives the world, corpus linguistics decontextualises the components of language. Corpus linguistics tends to look at language at the textual and lexical level. It is a mode of study that takes language out of its flow and reality, freezing it and rearranging it to give 'new perspectives on the familiar' (Hunston 2002, p. 3).[79] It draws on what is observable about language and how language is used to draw conclusions about how the author is

[78] A version of this chapter was published in a special edition of the *Journal of Gender & History: Gender Histories Across Epistemologies* (see Moreton 2012). I would like to formally thank the guest editors of this journal, as well as the peer review team, for their very generous and considered feedback and comments during the writing process. This publication marked the beginning of my PhD research into historical emigrant letter collections.
[79] The epistemological assumptions that underpin corpus linguistics as a methodology were also discussed by Professor Guy Cook at the 2011 Sinclair Open Lecture at the University of Birmingham, UK.

using language. The conclusions drawn are based on empirical data collection: frequencies, distributional patterns and proportions, and because of the design of the corpus, it is possible to move between the individual and the group and back again, noticing what is typical or unusual about one text when compared with many texts. Corpus linguistics, as applied to the study of correspondence, takes language out of context, reorganises it to notice new things based on quantitative investigation, and then puts the findings back into context to try to build a picture of the life and experiences of the author. This approach makes it possible to investigate systematically the language used by different authors and then to notice what those authors each have in common. As such, it provides a multi-layered approach to examining language and gender, allowing the analyst to test whether linguistic observations are about gender alone or gender in combination with other social, cultural or economic factors (such as age, class, location or level of education, for instance).

What is a corpus and what can it do?

As mentioned in footnote 16 of the introduction, a corpus can be defined as a ‘bod[y] of naturally occurring language data stored on computers’ (Baker 2006, p. 1). The ‘body of naturally occurring language’ can be anything from a few sentences to a large set of texts (the term ‘texts’ here refers to both written language and spoken transcriptions), but the main point to emphasise is that the data has been collected for a specific purpose, with the aim being something ‘other than to preserve the texts themselves because they have intrinsic value’, which is, as Hunston explains, what distinguishes a corpus from a digital archive (2002, p. 2). A corpus does not simply preserve and store texts so that they can be accessed more easily and by a wider number of people; rather, a corpus is designed with the intention of being representative of a particular type of text – newspapers, academic essays, letters, political speeches and so on, from a particular era, on a particular subject or by a particular socio-economic group and so on. Representativeness is usually achieved by 'breaking the whole down into component parts and aiming to include equal amounts of data from each of the parts' (Hunston 2002, p. 28). So, for example, a corpus of political speeches during the UK general election campaign of 2015 might include an equal weighting of speeches for the news media, TV debates and public addresses, by a range of politicians from the various parties. What goes into the corpus, then, depends on what the corpus will be used for and what research questions it will seek to address. However, as Hunston explains, it will also depend on 'what [data] is available', and quite often the analyst is negotiating a fine balance between selecting texts that are representative and working with whatever texts are available (2002, p.26). This issue of representativeness is always problematic and arguably no more so than when working with letters. In the case of emigrant letters, the analyst is always working with what is available – designing a representative corpus of personal letters is simply not achievable as there is no way of accounting for the experiences of those emigrants who chose not to write, who could not write or whose letters were lost, destroyed, or, years later, for various reasons, not donated. However, representativeness, in practice, is always a matter of degree, where fully representative material is not a possibility. A small, or limited, corpus can be regarded as sufficiently representative for a sufficiently delimited task, for instance.

The second thing that distinguishes a corpus from a digital archive is the way in which data content is explored and analysed. Although the data in a digital archive may be accessed online, without the need physically to visit a library or an archive, the content is generally studied linearly (as one would do with an original manuscript). Digitisation alone (and by that I mean optical character recognition (OCR) scanning or transcribing the letters and saving them in an electronic format) makes a document more accessible and to a certain extent more searchable (in a very limited sense of the term); however, it does not allow the collection to be explored in depth, or in creative ways. With a corpus, the data is stored in such a way that ‘it can be studied non-linearly, and both quantitatively and qualitatively’, using computer software (Hunston 2002, p. 2). The data (in this case the emigrant letter) can be marked-up in various ways – for contextual information such as gender, age, date of correspondence, socio-economic status, religious denomination, location (home and New World); for key themes such as homesickness, work, family, health, or for pragmatic features such as apologising, making requests, humour and so on. The data can also be annotated for parts of speech (word classification) and semantic categorisation. This markup and annotation allows individual letters and subgroups to be easily searched and compared in relation to one another and in relation to the whole. Additionally, computer software allows the content of the corpus, or subsections of the corpus (known as subcorpora) to be explored in ways that would be difficult, and in many cases impossible (depending on the amount of data being examined and the type of search being carried out), using more traditional methods of content analysis. Computer software allows the analyst to observe recurrent patterns, distributional trends and other statistical features, which would be hard to notice through reading alone. For this reason, it is often the data that will lead the investigation, pointing the analyst to features of the texts which they may not have noticed otherwise.

The starting point

The starting point for a corpus investigation is quantitative. What is unusual, interesting or typical about a text can only be explained by comparing it against other texts. It is this ‘comparative information that quantitative corpus data can provide’ (Stubbs 2008, pp. 230-43). By constantly moving between the cohort and the individual it is possible to notice both what is typical and distinguishing about a text or texts. In this chapter I will investigate the letters of the four Lough sisters who emigrated from Ireland to the United States in the mid to late nineteenth century.[80] I will examine the collection (or corpus) as a whole to see if there are any recurring patterns or phraseology, and what these might reveal; I will also compare the letter series (or subcorpora) of each individual sister to see how their language differs, and what this might reveal. I will, where relevant, use two reference corpora of emigrant correspondence from around the same period: twenty-one randomly selected letters by male Irish authors from a range of socio-economic backgrounds and twenty-one letters by female Irish authors. Letters for the reference corpora were borrowed (and transcribed) from Professor Miller’s archive. These two corpora (although very small for the purpose of this study) will allow me to test whether the findings from the *LOUGH Corpus* are

[80] The complete letter collection is described as the *LOUGH Corpus*. Throughout this thesis the individual letter series (subcorpora) for each sister are described as: *ELC* (which contains letters written by Elizabeth Lough), *NLC* (which contains letters by Annie, also referred to as Nan, Lough), *ALC* (which contains letters by Alice Lough) and, finally, *JLC* (which contains letters by Julia Lough). All italicised words and phrases are examples taken from the letters. Words in capitals represent the lemma (that is all variations of a particular word form, so BE would represent all forms of the verb to be: is, am, are, was, were etc.). Raw frequencies are presented in angle brackets.

representative of female emigrant correspondence more generally as well as the extent to which the language of male and female authors differs.

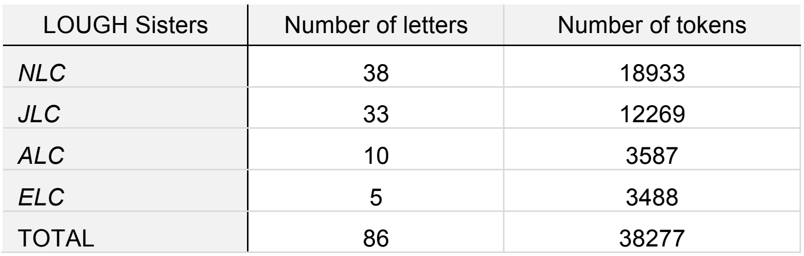
As mentioned in chapter one, there is a total of ninety-nine letters in the Lough collection held at the University of Missouri; however ten of these were excluded from this investigation as they did not contain sender information, making it difficult to assign these letters confidently to one of the four subcorpora – especially in instances where the original manuscript (or photocopy of that manuscript) is no longer available. As I will be comparing the letter series of each individual sister it is important that each correspondence is correctly assigned to a subcorpus – a wrongly assigned letter could affect the results. Another three letters were discounted, as these – although part of the Lough collection – were not written by any of the four sisters. Although the corpus is relatively small (compared with many corpora, for example, the *British National Corpus*[81] or the *Bank of English*,[82] which reach into millions of words), it will nonetheless provide a good foundation on which later studies, looking at larger bodies of data, can build. Corpus linguistics is about making comparisons by looking at what happens in one text and then seeing if this is typical of many texts, and vice versa. The same statistical measures are used when looking at a small amount of data as when looking at a large amount, thereby making it possible to compare corpora or subcorpora of different sizes.

To prepare the letters for corpus analysis they first needed to be digitised and then saved in plain text format (as mentioned in chapter one, this format is

[81] *The British National Corpus*, version 3 (BNC XML Edition) (2007) Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available from: http://www.natcorp.ox.ac.uk/.

[82] *The Bank of English* (1991 [2002]) COBUILD and The University of Birmingham. Available from: http://www.titania.bham.ac.uk/docs/svenguide.html.

compatible with most corpus software programmes). The process of digitisation and markup and the issues and challenges of working with original manuscripts and various versions of transcriptions will be discussed later in this thesis. It was not necessary to mark-up the letters for contextual information at this stage (date, location, etc.); nor was it necessary to annotate the letters for parts of speech as this study is intended to be data-led (i.e. basic frequency information will lead the investigation; I will not be approaching the corpus with specific, predetermined grammatical/structural searches in mind). All quantitative findings will need to be examined qualitatively (using concordance lines, which display the words in context) to establish how a word or phrase is functioning – whether as a noun, verb, adjective, etc.



LOUGH Sisters	Number of letters	Number of tokens
NLC	38	18933
JLC	33	12269
ALC	10	3587
ELC	5	3488
TOTAL	86	38277

Table 2.1: The *LOUGH Corpus*

As shown in Table 2.1, (after removing those letters which cannot be assigned to one of the Lough sisters) the *LOUGH Corpus* contains eighty-six letters – a total of 38,277 words. Annie Lough, the third sister to emigrate in 1878, appears to have written the most letters of the four sisters – a total of thirty-eight letters (18,933 words) between 1890 and 1928, nine of which were to her mother and twenty-six to her sister Mary (see Table 2.2), both of whom remained in the Lough's home town – Meelick, Queen's County, Ireland – until their deaths. Additionally, there is one letter addressed both to Annie's mother and sister and a further two letters to her nice and nephew. Julia Lough, the last sister to emigrate in 1884, also wrote regularly – mainly to her mother (twenty-three letters) and also her sister (ten letters) – a total of thirty-three letters (12,269 words) between 1884 and 1927. Elizabeth and Alice were the first sisters to emigrate between 1870 and 1871, yet they wrote the smallest number of letters. Elizabeth wrote five letters (3,488 words) to her mother, father and sisters between 1876 and 1877, when she first emigrated to the US and Alice wrote ten letters (3,587 words) to her sister and mother between 1888 and 1914 (two when she first emigrated and then another three at roughly five-year intervals – five of the letters are not dated, but the content would suggest they were written several years after emigrating). It should be pointed out, however, that this information is based on the number of letters held in Professor Miller's archive (in other words, the number of letters which were donated). As mentioned previously, when discussing the issue of representativeness, there is no way to know how many letters were actually sent or how many were lost or destroyed.

	No. of letters sent			
Addressee	*NLC*	*JLC*	*ELC*	*ALC*
Mother	9	23	1[83]	3
Sister (Mary Lough, later Fitzpatrick)	26	10	0	7
Mother and Sister	1	0	0	0
Nephew - James	1	0	0	0
Niece - Alice	1	0	0	0
Father, Mother and Sisters	0	0	2	0
Father and Mother	0	0	2	0

Table 2.2: Breakdown of senders/recipients

[83] Miller believes this is a fragment belonging to another of Lizzie's letters which is addressed to her mother and father. For now, however, I have recorded this as a separate letter.

Having grouped the Lough data, it was then possible to explore the content of the letters using computer software. There are a number of useful corpus analysis programmes available, some of which are web-based – *Wmatrix*[84] and *Sketch Engine*[85] – while others are computer-based – *AntConc*,[86] *WordSmith*[87] and *ConcGram*.[88] I have chosen to use *AntConc* for two reasons: first, it is freely available online and second, it has certain functionalities which I am interested in using for this investigation – specifically, the n-gram procedure which will be discussed later in the chapter.

Simple frequency data

The first calculation that *AntConc* can provide is something called a type/token ratio, which can be obtained for the corpus as a whole and for each subcorpus. The term 'token' refers to the total number of words in a corpus. The term 'type' refers to the number of original (or different) words in the corpus. So, for example, the word HOME occurs 193 times in the *LOUGH Corpus*, which would equal 193 tokens, but would only constitute one type.[89] Types are visibly distinct forms, so that while many might want to treat 'home' and its plural 'homes' as one lemma and one word, they are two distinct types in this mechanical calculation. Likewise, this calculation will not distinguish rather different

[84] Rayson, P. (2009) *Wmatrix*. Lancaster University. Available from: http://ucrel.lancs.ac.uk/wmatrix/.

[85] Kilgarriff, A. and Kosem, I. (2012) Corpus Tools for Lexicographers. In S. Granger and M. Paquot (eds.), *Electronic Lexicography*. New York: Oxford University Press. Pp. 31-56. Available from: http://www.sketchengine.co.uk.

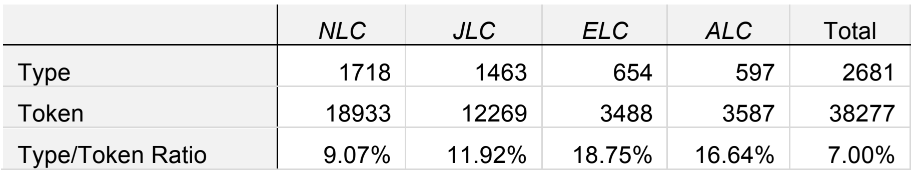
[86] Anthony, L. (2011) *AntConc* Version 3.2.2 [Macintosh OS X]. Tokyo, Japan: Waseda University. Available from: http://www.laurenceanthony.net/.

[87] Scott, M. (2004) *WordSmith Tools Version 4*. Oxford: Oxford University Press. Available from: http://www.lexically.net/wordsmith/index.html.

[88] Greaves, G. (2005) *ConcGram*. Amsterdam: John Benjamins Publishing Company. Available in CD-ROM from: https://benjamins.com/#catalog/software/cls.1/main.

[89] Note that *AntConc* does not distinguish between word class (unless the data is tagged for Parts of Speech), so HOME, whether it is used as a noun or an adjective, would be categorised as one 'type'.

meanings, as in ‘home in on’ versus ‘a good home’. The type/token ratio is calculated by dividing the number of types by the number of tokens – this figure is then expressed as a percentage. A low type/token ratio (i.e. many tokens of a small number of types, yielding a low percentage) suggests that certain words are being used over and over again. A high type/token ratio (i.e. many types, with few tokens for each type, tending towards a higher percentage) suggests a more diverse range of language is being utilised, with fewer words being repeated. Looking at Table 2.3, the data shows that the *LOUGH Corpus* has an overall type/token ratio of 7%. Breaking this down by subcorpora, the data shows that Annie has the lowest type/token ratio (9.07%), followed by Julia (11.92%). The type/token ratios for Alice and Lizzie are slightly higher, 16.64% and 18.75% respectively.



	NLC	*JLC*	*ELC*	*ALC*	Total
Type	1718	1463	654	597	2681
Token	18933	12269	3488	3587	38277
Type/Token Ratio	9.07%	11.92%	18.75%	16.64%	7.00%

Table 2.3: Type/token ratios for the *LOUGH Corpus*

On the surface, this might suggest that Annie’s and Julia’s letters are more formulaic and repetitive, whereas Alice’s and Lizze’s letters contain greater lexical variety and complexity. However, it is more likely that this difference in percentages reflects the size of the corpora. The larger the corpus the more likely some words, particularly grammar words, are repeated, which in turn will reduce the type/token ratio. To demonstrate this, Table 2.4 and Figure 2.1 show the accumulative type/token ratios, year after year, for each sister (note that only the letters containing a date are included in this investigation). Taking Julia as an example, the data shows that her first letter in 1884 contains sixty-one tokens (or words) and forty-six types, giving an overall type/token ratio of 75.41%, suggesting that the letter contains a good amount of linguistic diversity with relatively little repetition (although not surprisingly since this is a very short letter). However, looking at the accumulative figures for letters sent between 1884 and 1889 (1,933 tokens and 530 types), there is a much lower type/token ratio (27.42%), which would suggest that some repetition is occurring in the five letters sent during this period. This is to be expected: as mentioned earlier, the larger the corpus the more likely it is that words will be repeated; however, the formulaic nature of letter writing perhaps also goes some way to explaining the dramatic drop in type/token ratio from Julia's first letter sent in 1884, which has a type/token ratio of 75.41%, to her last letter sent in 1927, which has a type/token ratio of 13.89%. The extent of the formulaic writing would require further investigation. Certainly, the openings and closing are likely to follow a standard format, but it would be interesting to examine the body of the letters to see whether they too adopt a set pattern, with less new information being presented over time.

Toolan suggests that what are potentially very interesting when examining accumulative type/token ratios in extended narratives are any sharp 'spikes' or 'dips' in the predictable decline in type/token ratios (2009). In the Lough data there is a sharp decline, or dip, between Julia's first letter sent in 1884 (75.4%) and her second letter sent later that year (41.7%) – a difference of 33.7%, meaning that Julia's second letter is covering a lot of the same lexical ground (and perhaps the same topics) as her first did. Similarly for Annie, there is

JULIA																								
Letter No.	5	6	8	9	10	13	15	17	18	19	20	21	23	26	27	77	30	28	29	32	58			
Date	1884	1884	1888	1889	1889	1890	1890	1890	1891	1891	1891	1891	1892	1893	1893	1893	1893	1894	1894	1895	1927			
Accumulative Type	46	250	335	458	530	592	641	674	724	768	791	823	878	914	942	983	1001	1083	1121	1160	1198			
Accumulative Token	61	600	977	1430	1933	2429	2814	3188	3550	3907	4243	4562	5004	5356	5828	6192	6545	7310	7810	8251	8622			
Ratio %	75,41	41,67	34,29	32,03	27,42	24,37	22,78	21,14	20,39	19,66	18,64	18,04	17,55	17,06	16,16	15,88	15,29	14,82	14,35	14,06	13,89			
ANNIE																								
Letter No.	12	70	22	25	31	33	34	37	38	40	42	43	44	36	46	48	49	52	53	55	56	57	59	60
Date	1890	1891	1891	1893	1895	1898	1899	1901	1901	1902	1906	1906	1906	1910	1912	1913	1914	1918	1918	1919	1919	1925	1928	1928
Accumulative Type	205	417	477	611	653	653	721	753	792	817	848	870	891	975	1039	1065	1087	1149	1251	1326	1396	1380	1389	1424
Accumulative Token	491	1532	2027	2965	3580	4621	5168	5571	6019	6393	6833	7174	7479	8005	8651	9158	9565	10224	11104	11932	12444	12688	12901	13536
Ratio %	41,75	27,22	23,53	20,61	18,24	14,13	13,95	13,52	13,16	12,78	12,41	12,13	11,91	12,18	12,01	11,63	11,36	11,24	11,27	11,11	11,22	10,88	10,77	10,52
ALICE																								
Letter No.	7	11	41	45	50																			
Date	1888	1889	1904	1910	1914																			
Accumulative Type	170	305	392	436	464																			
Accumulative Token	395	928	1380	1756	2042																			
Ratio %	43,04	32,87	28,41	24,83	22,72																			
LIZZIE																								
Letter No.	1	2	3	4																				
Date	1876	1876	1876	1877																				
Accumulative Type	409	502	543	631																				
Accumulative Token	1450	2033	2390	2939																				
Ratio %	28,21	24,69	22,72	21,47																				

Table 2.4: Accumulative type/token ratios for each subcorpus

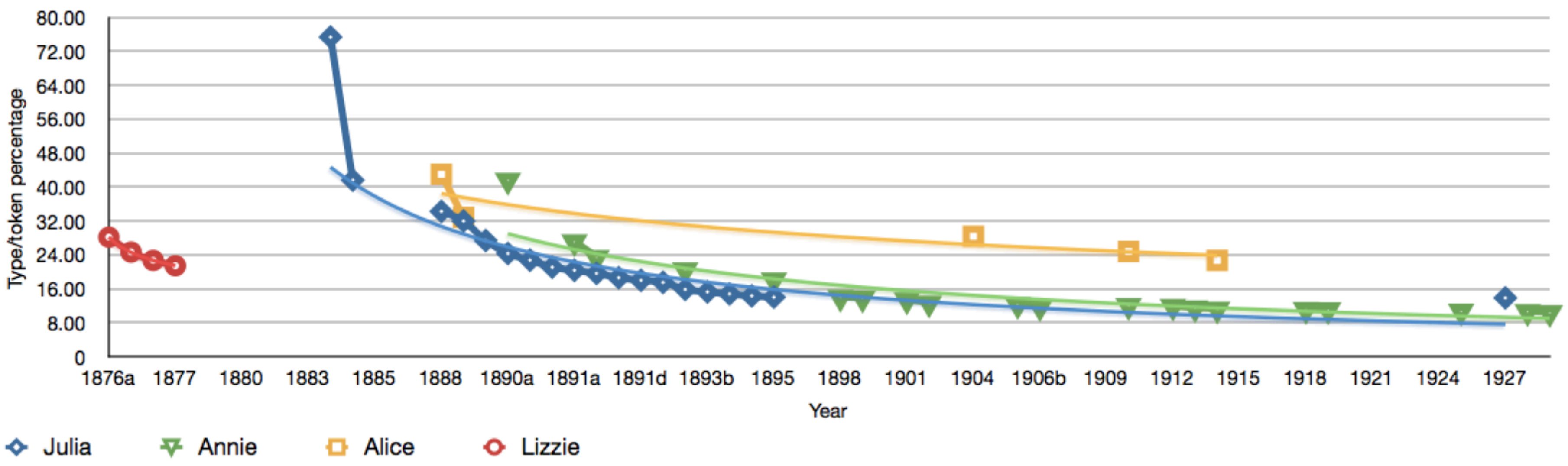
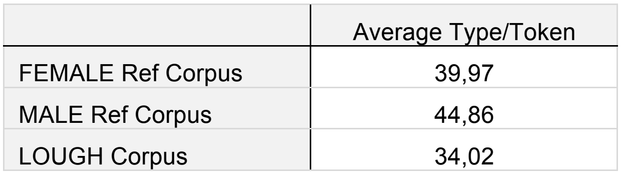


Figure 2.1: Accumulative type/token ratios for each subcorpus

a noticeable difference of 14.6% between her first letter sent in 1890 (41.8%) and her second letter sent in 1891 (27.2%), after which the decline is much less pronounced. These dips could be explained, in part, by the length of the letters: in both cases the first letter is quite short whereas the second letter is much longer; however, it could also be indicative of a more formulaic writing style adopted by the two younger sisters, possibly indicative of differences in education between the Lough sisters, although further analysis would be needed to test this hypothesis.

An average type/token ratio can also be calculated. This goes some way to resolving the problem of type/token ratios being lower for larger corpora and higher for smaller corpora, and allows data sets of different sizes to be compared. This is done by calculating the type/token ratio for the first 1,000 words of a corpus, then the next 1,000 words, then the next, and so on. Finally, an average is calculated based on these figures. Table 2.5 shows the average type/token ratio for the *LOUGH Corpus* and the two reference corpora (*FEMALE Ref.* and *MALE Ref.*). The data shows that female authors have a slightly lower type/token ratio (39.97%) compared with male authors (44.86%). The average type/token ratio for the *LOUGH Corpus* is slightly lower than both reference corpora at 34.02%, which might support earlier observations that the Lough letters (particularly in the case of Julia Lough and Annie Lough) are perhaps more formulaic than one might expect – an observation that is certainly worth further investigation: what is being repeated and what function does this repetition serve?

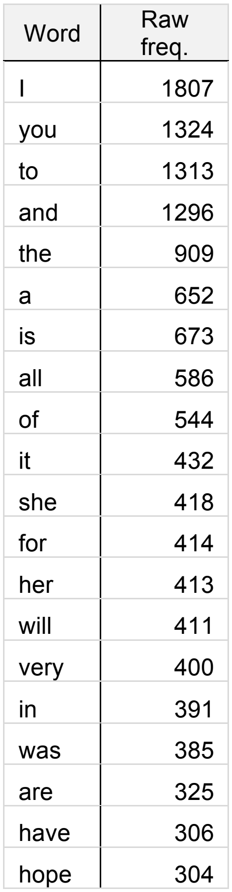


	Average Type/Token
FEMALE Ref Corpus	39,97
MALE Ref Corpus	44,86
LOUGH Corpus	34,02

Table 2.5: Average type/token ratios across three corpora

Words and frequencies

Having established the lexical density of the letters, the next stage is to look at which words are being repeated (or not, as the case may be). Using *AntConc* it is possible to create wordlists for the whole corpus and each subcorpus. Table 2.6 shows the twenty most frequently occurring words in the *LOUGH Corpus*.



Word	Raw freq.
I	1807
you	1324
to	1313
and	1296
the	909
a	652
is	673
all	586
of	544
it	432
she	418
for	414
her	413
will	411
very	400
in	391
was	385
are	325
have	306
hope	304

Table 2.6: Wordlist for the *LOUGH Corpus*

The left column ('Word') shows the words listed in order of frequency with the right column (Raw freq.) providing the actual number of occurrences. Table 2.6 shows that grammar words are most common: *I* <freq. 1,807>, *you* <freq. 1,324>, *to* <freq. 1,313>, *and* <freq. 1,296>, *the* <freq. 909>. Grammar words are the glue that holds the content together – so it is perhaps not surprising that these words occur more frequently. However, the propensity for certain grammar words over others can be equally revealing. Table 2.6, for example, shows that the pronouns *I* and *you* are the most frequently occurring words in the *LOUGH Corpus* with *I* scoring slightly higher than *you*: <freq. 1,807> versus <freq. 1,324> (a ratio of 4:3). One might expect the first person singular pronoun *I* to score high in ego-documents such as letters; however, previous studies have identified gendered variations in terms of pronoun usage. McLelland (2007), for example, found that female authors tended to refer to themselves using the first person singular pronoun *I* more than male authors; and a study by Nurmi and Palander-Collin (2008) found that pronoun usage reflected the power relations between author and recipient – when the relationship was equal (letters between siblings, for example) the first person pronoun usage was high; when the relationship was unequal (letters between children and parents, for example) the first person pronoun usage was low. Their study also found that the sex of the recipient had an effect on pronoun usage, with authors referring to themselves more frequently using *I* when the recipient was female. The current study supports some of these findings with *I* occurring more frequently in the *LOUGH Corpus* (an average of 47.21 occurrences per 1,000 words) and the FEMALE Corpus (an average of 41.90) than in the MALE Corpus (an average of just 32.94) – see Table 2.7. The findings did not, however, support Nurmi's and Palander-Collins's observation that first person pronoun usage tends to be greater in letters between authors and recipients of equal status (such as siblings); instead, in the Lough letters, the data showed that *I* occurs slightly more frequently in letters addressed to the mother (an average of 50.37 occurrences per 1,000 words) than in those addressed to the sister (an average of 45.29 occurrences) – see Table 2.8. Note that the 'normalised' figures in Table 2.7 and Table 2.8 allow meaningful comparisons to be made across data sets of different sizes. It is calculated by dividing the raw frequency by the number of tokens x 1,000. This gives an average frequency (of a particular word or phrase) per 1,000 words.

	I (Raw freq.)	*I* (Normalised)	*You* (Raw freq.)	*You* (Normalised)
LOUGH Corpus	1807	47.21	1324	34.59
FEMALE Ref.	693	41.90	353	21.34
MALE Ref.	681	32.94	291	14.07

Table 2.7: Occurrences of *I* in each corpus

No. of letters to:	*I* (Raw freq.)	*I* (Normalised)	*You* (Raw freq.)	*You* (Normalised)
Sister (48)	919	45.29	772	38.04
Parents (43)	957	50.37	592	31.16

Table 2.8: Occurrences of *I* in letters sent to parents/siblings[90]

The pronoun *you* is also potentially very interesting as it has the ability to occupy two grammatical positions (Subject and Object), so its usage might reveal something about the author/recipient relationship: how the authors are positioning themselves and how they are positioning the recipient. Analysis of the concordance lines for *you* shows that it frequently occurs in the position of Subject of what can be described as a projected clause, where the projecting

[90] Note: all letters were included in this investigation (including letters where the authorship is unknown), provided the letter was specifically addressed to either 'mother', 'mother/father' or 'sister'.

clause contains the pronoun *I*, as in ***I** hope* **you** *will write soon* (*I hope* being the projecting clause: the part which projects an idea, fact or proposition; and *you will write soon* being the projected clause: the idea, fact or proposition that is being projected). In short, *you*, in these occurrences, is the real or psychological Subject of these sentences. Additionally, there is an argument, taking this further, that these sentences show underlying deontic modality: *You should write (to me) soon*.[91] This is down-toned and made indirect by the double boulomaic or willingness modalities, since *I hope you* can be paraphrased as *I would wish that you would want*.

Table 2.9 shows that the pronoun *I* (in these projection clauses) most commonly occurs either two words to the left (L2) <freq. 147> or three words to the left (L3) <freq. 27> of the search word *you*, with the most frequent structures being *I hope you* <freq. 95>, *I suppose you* <freq. 28>, *I am sure you* <freq. 15>, *I wish you* <freq. 11>, *I know/no you* <freq. 11>, *I think you* <freq. 7> and *I am glad you* <freq. 4>). In these instances, the projecting clause (i.e. the clause which introduces the projected clause – the main fact, idea or proposition) contains a mental verb or an adjective carrying epistemic modality (such as *suppose* or *sure* (expressing probability or certainty)), or a mental verb carrying boulomaic modality (such as *hope* or *wish* (expressing desire or volition)). It is, arguably, at this point that a phraseological pattern begins to emerge: *I* + *Verb* + *You*; *I* + *BE* + *Adj* + *You*. In any case, the prominence of *you* as doer, agent or focalised, constructed centre of attention, is very striking. I will talk more about projection clauses later in the chapter.

[91] Deontic modality indicates 'the necessity…of the proposition in the utterance' (Jeffries and McIntyre 2010, p. 78).

No. of words to the left	L7	L6	L5	L4	L3	L2	L1	**NODE**
Freq. of ‘I’	0	3	4	6	27	147	0	**YOU**

Table 2.9: Position of *I* in projecting clauses

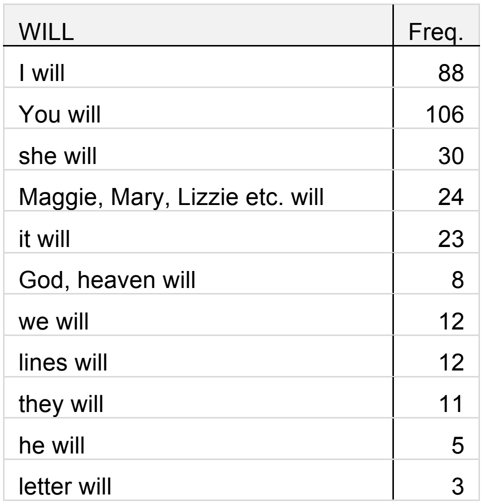
Breaking the wordlist down further, Table 2.10 provides the word frequency lists for each subcorpus (as well as the corpus as a whole). The data shows that the grammar words *I* and *you* score high in all four subcorpora; however, although there are <759> occurrences of *I* in the *NLC* and only <236> occurrences in the *ELC*, statistically Lizzie is using *I* much more frequently than the other sisters – on average 67.66 times per 1,000 words, compared with 40.09 for Annie, 50.53 for Julia and 53.53 for Alice. Annie appears to be using *I* (40.09) and *you* (38.66) almost on a 1:1 ratio, perhaps suggesting that she is often directly involving or addressing the recipient in her letters; whereas Lizzie is using *I* (67.66) approximately two and a half times more frequently than she is using *you* (22.94), perhaps suggesting that her letters are more author focused. In all subcorpora the same grammar words (*I*, *you*, *and*, *to*, *the*) are being repeated, which may indicate that certain grammatical structures are also being repeated; this, in turn, may go some way to explaining the low type/token ratio discussed earlier, although further exploration would be needed before any conclusions could be drawn.

Another possible avenue for investigation is the use of *will*, which ranks high across three of the subcorpora: <freq. 229> in the *NLC*, <freq. 123> in the *JLC*, and <freq. 40> in the *ALC*. The modal verb *will* is interesting as it has several different functions and can be used to express epistemic modality (i.e. certainty/probability), or boulomaic modality (i.e. desire/volition). There are

LOUGH Corpus		*ALC*			*JCL*			*ELC*			*ALC*		
Raw freq.	Normalised	Word	Raw freq.	Normalised	Word	Raw freq.	Normalised	Word	Raw freq.	Normalised	Word	Raw freq.	Normalised
1807	47,21	and	767	40,51	I	620	50,53	I	236	67,66	I	192	53,53
1324	34,59	I	759	40,09	to	422	34,4	to	126	36,12	to	147	40,98
1313	34,30	you	732	38,66	and	391	31,87	you	80	22,94	you	124	34,57
1296	33,86	to	618	32,64	you	388	31,62	is	73	20,93	and	107	29,83
909	23,75	the	508	26,83	the	249	20,3	the	71	20,36	the	81	22,58
652	17,03	all	343	18,12	a	208	16,95	she	70	20,07	very	61	17,01
673	17,58	a	339	17,91	is	201	16,38	it	50	14,33	all	56	15,61
586	15,31	is	311	16,43	all	152	12,39	of	47	13,47	of	54	15,05
544	14,21	of	292	15,42	of	151	12,31	an	42	12,04	is	52	14,50
432	11,29	it	236	12,47	her	140	11,41	for	40	11,47	a	51	14,22
418	10,92	will	229	12,1	am	137	11,17	not	40	11,47	me	43	11,99
414	10,82	for	226	11,94	she	134	10,92	her	39	11,18	her	40	11,15
413	10,79	very	218	11,51	in	127	10,35	my	38	10,89	will	40	11,15
411	10,74	in	210	11,09	will	123	10,03	he	37	10,61	she	39	10,87
400	10,45	hope	201	10,62	for	122	9,94	all	35	10,03	they	39	10,87
391	10,22	was	201	10,62	not	116	9,45	was	35	10,03	was	38	10,59
385	10,06	her	194	10,25	have	113	9,21	but	32	9,17	have	37	10,32
325	8,49	are	184	9,72	it	113	9,21	have	32	9,17	in	34	9,48
306	7,99	she	175	9,24	was	111	9,05	no	32	9,17	it	33	9,20
304	7,96	well	157	8,29	are	109	8,88	am	31	8,89	but	32	8,92

Table 2.10: Wordlist for each subcorpus

<411> occurrences of *will* in the *LOUGH Corpus*. Six of these occurrences show *will* functioning as a noun (as in *God's will* and *holy will*), so these can be discounted, leaving <405> occurrences of *will* functioning as a modal verb. Of these <405> occurrences almost half come after the pronouns *I* <freq. 88> and *you* <freq. 106> (see Table 2.11).



WILL	Freq.
I will	88
You will	106
she will	30
Maggie, Mary, Lizzie etc. will	24
it will	23
God, heaven will	8
we will	12
lines will	12
they will	11
he will	5
letter will	3

Table 2.11: Occurrences of WILL in the *LOUGH Corpus*

Looking more closely at the concordance lines for *I will* the data shows that in most instances (<61> out of <88> occurrences) *will* is being used in 'signing off' structures to signal the close of the letter (see concordance lines below for examples), with the meaning being one of intention. All of the instances below – *I* ***will*** *conclude*, *I* ***will*** *finish*, *I* ***will*** *bring my letter to a close* – could be substituted with *I intend to* and as such are expressing boulomaic modality. As with '*I am writing to you because . . .* ' and '*You ask me X, so I will tell you . . .* ', these meta-discursive phrases help to structure the text as well as serving an interactive function.[92]

[92] For more information on metalanguage and metadiscursive phrases see Gee (2008) and Ädel (2006).

```
ope they are well so Dear mother I will bring my letter to a close I hope you
I know they are but very few now I will close now dear Sister with best and kindest
very well with it so Der Mother  I will conclude with fondest and best love to you
love to them and now dear Sister I will finish as I cannot wish you a merry Xma
ou know I am thinking of you and I will not forget you next year with Gods help.
yo will write as soon as you do. I will not write again till I get an answer to
when the last one was not a girl I will not say any more now till I hear from you
isitor as you I dont ask for any I will say good by now with love to you and an
ry you will write soon again and I will send you a longer letter next time and
see their grandfather some times I will try and send you their pictures some
I was going to write to Mary but I will wait now till I get her next letter y all
t I will have it before Xmas and I will write to you again before Xmas With the
```

Figure 2.2: Sample concordance lines for *I will*

Concordance lines display the search term (in this case *I will*) in context. The concordance lines are presented this way (that is with the search term centrally aligned) so as to allow the analyst to notice linguistic patterns – words that typically appear to the right or left of the search term.

The concordance lines for *you will* show that what follows is a limited range of verbs. There are verbs to do with the act of sending/receiving letters (*write*, *send*, *receive*, *get*); and there are verbs to do with cognition (*like*, *forgive*, *excuse*). (See concordance lines below. Note that the verb *keep* is difficult to categorise as it functions in very different ways and has different meanings depending on the context in which it is being used. In the concordance lines being examined here, *keep* is used in the context of *you will keep to your promise*, where *keep* is part of a fixed expression, meaning ‘fulfil your agreement’). In all of these instances of *you will* it is difficult to know whether *will* is expressing epistemic or boulomaic modality as there is not enough context for either function to predominate. When, for example, the author says, *you will forgive me for not writing before now*, it is not clear whether *will* is being used to express certainty (as in ‘I am quite sure that you will forgive me’), or desire/volition (as in ‘I want you to forgive me’ or ‘I hope you intend to forgive me’).

```
you will              forgive me for not writing before now my son
you will              send the paper you promised me I
you will              excuse all mistakes
you will      →       soon write Dearest Mother love to you May john
you will              keep to your promise and write again to me
you will              get them Dear Mother I will finish up for this
you will              receive in due time. the censors are kept very
you will             like to reed them let me know if the letters
```

Figure 2.3: Sample concordance lines for *you will*

An investigation of the wider context, however, reveals that in most cases (<90> out of <106> occurrences) *you* is the Subject and *will* is the auxiliary modal of a projected clause, preceded by a projecting clause (see Figure 2.4), with the most frequent patterns being *I hope you will*, *I suppose you will* and *I am sure you will*.

With this wider phraseological context it now becomes more possible to determine the function of *will* in these instances. The type of modality (whether epistemic or boulomaic) is projected onto the recipient via the projecting clause, pushing a mild obligation, or placing social pressure onto the addressee to respond in a certain way. The concordance lines for *you will* seem to suggest that *will* is more frequently used to express boulomaic modality, with the main pattern (*I hope you will* + *V*) being used to express the author's desire for the recipient's willingness to do something. Through these clauses the author's wants, needs, desires or intentions are transferred onto the recipient – they become the recipient's own and create a psychological bond between both participants.

Another observation which can be made from Table 2.10 is that across all four subcorpora the only lexical word which appears (in the top twenty) is the verb *hope* with a frequency of <201> in the *NLC*. However, moving further down the wordlists more content words begin to appear. Table 2.12 provides a list of the ten most frequent lexical verbs in each subcorpus (i.e. the first ten lexical

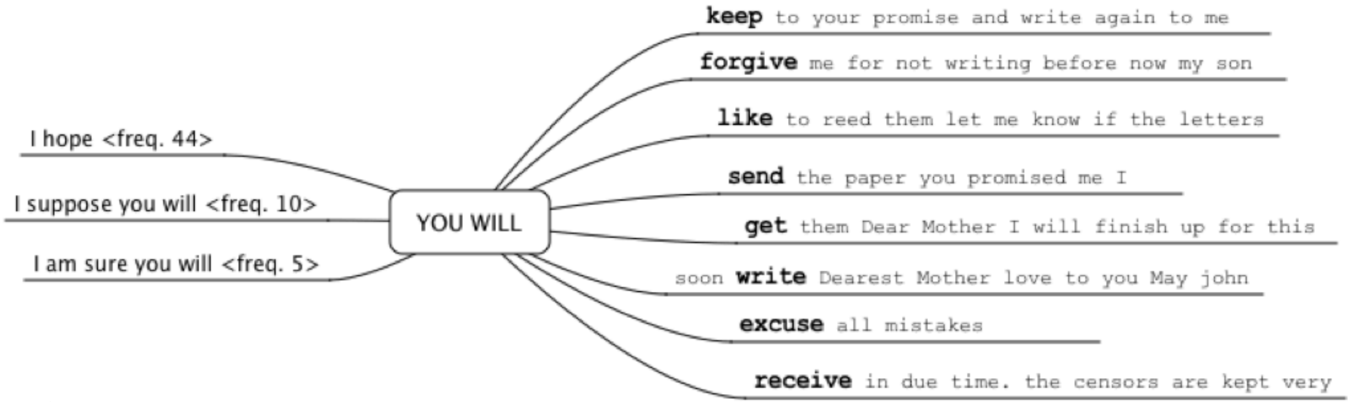

I hope <freq. 44>
I suppose you will <freq. 10>
I am sure you will <freq. 5>
YOU WILL
keep to your promise and write again to me
forgive me for not writing before now my son
like to reed them let me know if the letters
send the paper you promised me I
get them Dear Mother I will finish up for this
soon write Dearest Mother love to you May john
excuse all mistakes
receive in due time. the censors are kept very

Figure 2.4: Patterns for *you will*

LOUGH Corpus			ANNIE Corpus			JULIA Corpus			LIZZIE Corpus			ALICE Corpus		
Verb	Freq.	Norm.	Verb	Freq.	Norm.	Verb	Freq.	Norm.	Verb	Freq.	Norm.	Verb	Freq.	Norm.
hope	304	7,94	hope	201	10,62	hope	74	6,03	think	31	8,89	see	23	6,41
write	197	5,15	write	120	6,34	get	66	5,38	get	18	5,16	hope	19	5,30
get	151	3,94	see	74	3,91	hear	65	5,30	write	22	6,31	send	18	5,02
see	143	3,74	love	69	3,64	write	52	4,24	got	11	3,15	get	15	4,18
think	140	3,66	suppose	64	3,38	think	49	3,99	seen	11	3,15	like	15	4,18
love	127	3,32	know	57	3,01	know	42	3,42	thought	11	3,15	love	15	4,18
hear	112	2,93	get	52	2,75	see	40	3,26	hope	10	2,87	know	13	3,62
know	112	2,93	work	48	2,54	love	35	2,85	going	9	2,58	live	13	3,62
suppose	103	2,69	hear	47	2,48	go	34	2,77	love	8	2,29	think	13	3,62
work	84	2,19	think	47	2,48	let	25	2,04	taken	8	2,29	write	13	3,62

Table 2.12: Most frequent lexical verbs by subcorpus

verbs as they appear, in whatever form, in the wordlist). Note that I am looking at lexical verbs and not auxiliary verbs (HAVE, BE, DO), which tend to serve a grammatical function.

Looking at Table 2.12 there are two things that stand out. First, there are a high number of mental verbs of cognition, perception and desire (*hope*, *suppose*, *see*, *think*, *like*, *love*, *hear*, *know*), with certain verbs (*hope*, *love* and *think*) appearing across all four subcorpora. Second, nearly all of these verbs appear to be in their base form, with the following exceptions: Lizzie uses the past tense *got* and *thought*, the participle *seen* and *taken* and the continuous form *going*. The high frequency of base forms may in part be explained by the high frequency of *to* and *will* across all four subcorpora (as what tends to follow both *to* and *will* is the base form of the verb, as in *to hear*, *to think*, *will send*, *will go*). However, a closer look at the context surrounding these mental verbs in their base form (of which there are <1,115> occurrences) reveals that they are rarely used after *will* (just <25> instances); they are more frequently used after *to* (a total of <173> instances – the most common structures being *to see* <freq. 72> and *to hear* <freq. 75>); but they are most frequently used in the present tense after the first person singular pronoun *I* (a significant <453> instances).

The high frequency of these mental verbs of cognition, perception and desire is interesting for two main reasons. The first is that these verbs, as explained by Halliday and Matthiessen, ‘relate to inner experience (what we experience as going on inside ourselves, in the world of consciousness)’ and usually describe emotions, thoughts, or perceptions, thereby providing insight into the psychological worldview of the author (2004, p. 170). The second is that these verbs are special because they have the ability to project: that is they have the ‘ability to set up another clause “outside” the “mental” clause as the representation of the “content” of consciousness’ (Halliday and Matthiessen 2004, p. 206). This latter point appears to support previous findings which show a high frequency of the projecting clause *I hope* (*hope* being a mental verb of desire).

Halliday and Matthiessen make a distinction between the projection of propositions and the projection of proposals, with each type of projection having its own lexicogrammar. ‘Whereas propositions, which are exchanges of information [i.e. exchanges which require a verbal response], are projected mentally by processes [verbs] of cognition – thinking, knowing, understanding, wondering, etc. – proposals, which are exchanges of goods-&-services [i.e. exchanges which require a non-verbal response], are projected mentally by processes [verbs] of desire’ (2004, p. 461). Further, what is interesting about the lexicogrammar of proposals is that they can be followed by a future declarative (will + base form) or non-finite (including to-infinitive) dependent clause (as in *I hope you* ***will write*** *soon* or *I hope* ***to hear*** *from you soon*). So, when a verbal response is required the verb is likely to be one of cognition (as in *I know you are trying to do the best you can*). When a non-verbal response is required the verb is likely to be one of desire (as in *I hope you will write often*) – see Table 2.13.

Type of Exchange	
Proposition	**Proposal**
Exchange of information	Exchange of goods & services
Verbal response	Non-verbal response
Mental verbs of cognition (*know*, *think*, *suppose* etc.)	Mental verbs of desire (*hope*, *wish*, *want* etc.)
I know you are trying to do the best you can	*I wish you would write oftener*
I think you are growing smarter	*I hope you will send me the paper*
I suppose you are always busy	*I want you to give five shillings of mine to Mary*

Table 2.13: Examples of propositions and proposals

Words in context: n-grams and clusters

Some of the observations discussed so far begin to piece together when the next test is carried out, which looks at n-grams. N-grams are: X number of words which appear consecutively Y number of times. The analyst can set the parameters, so, for example, using the n-gram function within *AntConc* the analyst could search for all 3-grams (three words appearing consecutively) which occur five times or more in the corpus. Table 2.14 gives a summary of the most frequently occurring 2, 3, 4, 5 and 6-grams.

Table 2.14 shows that the quantitative findings discussed in previous sections are partly realised in these n-grams, with the lexical verb *hope* followed by the modal auxiliary verb *will* ranking high across four of the five searches. However, this only reveals part of the picture – there are <274> instances of the 2-gram *I hope*, but only <58> instances of *I hope you will*. To get a fuller understanding of the phraseology surrounding a particular word or phrase *AntConc* has the capability to search for clusters. The word tree in Figure 2.5 shows the three, four, five and six word clusters surrounding the phrase *I hope*. Figure 2.5 highlights the lexical and grammatical (or lexicogrammatical)

2-gram	Freq.	3-gram	Freq.	4-gram	Freq.	5-gram	Freq.	6-gram	Freq.
I am	284	**I hope you**	95	**I hope you will**	58	few lines will find you	12	these few lines will find you	10
I hope	274	**hope you will**	61	love to you and	22	**I hope you will write**	11	**I hope you will write soon**	9
and I	149	I am sure	56	I am sending you	21	with love to you and	11	**hope you will write soon and**	8
to you	147	and I hope	45	and I hope you	19	**and I hope you will**	10	and let me know all the	7
you will	121	Dear Mother I	45	and let me know	17	these few lines will find	10	let me know all the newes	7
all the	120	love to you	42	give my love to	16	and let me know all	9	love to you and John and	7
hope you	114	I am glad	41	Dear Mother I am	15	**hope you will write soon**	9	few lines will find you all	6
I was	108	let me know	40	I am sure you	15	I am sending you some	8	**hope these few lines will find**	6
to her	108	are all well	35	to hear from you	15	let me know all the	8	**I hope these few lines will**	6

Table 2.14: Most frequent n-grams in the *LOUGH Corpus*

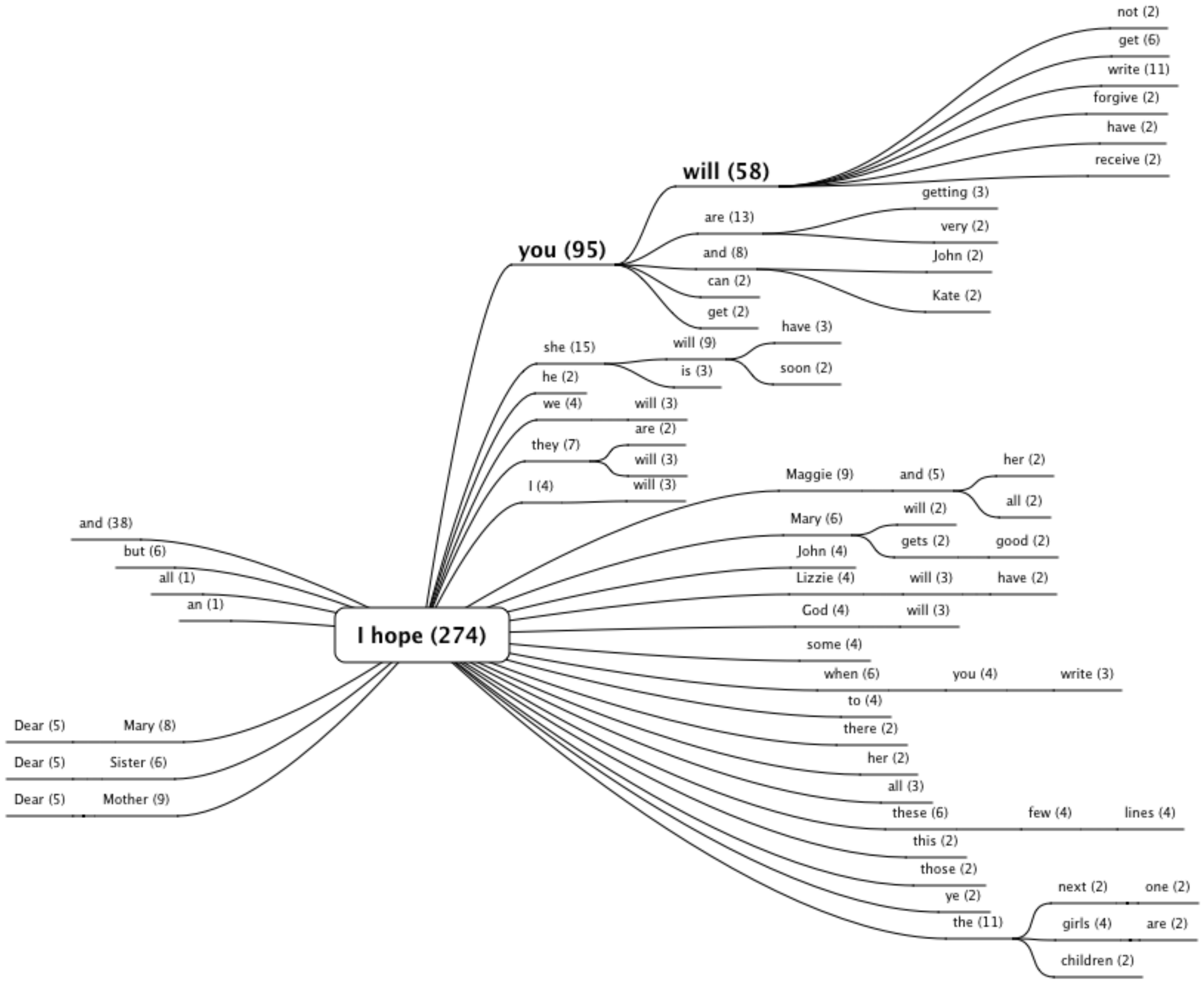


Figure 2.5: Three, four, five and six-word clusters for *I hope*

patterns surrounding *I hope*.[93] The diagram shows that not all options are equally probable. What most frequently follows *I hope*, in the *LOUGH corpus*, is the pronoun *you*; and what most frequently follows *I hope you* is the modal auxiliary *will*. As the tree branches out the lexicogrammatical choices become fewer, so *I hope you will not*, for example, occurs just twice in the corpus and in itself is not overly significant. Looking at the broader picture, however, through examining lots of evidence at the same time, a phraseological pattern for *I hope* begins to emerge. The question then would be whether this pattern is typical of this data set only, or typical of letters/personal narratives more generally? Is this phraseology used more by one author than another? Finally, what do these linguistic choices reveal about the author, their sex or their experiences?

The quantitative observations so far have teased out several possible lines of inquiry, which could be examined qualitatively. The analyst might, for example, investigate the low type/token ratio and whether or not the significant dip in ratio between Julia's and Annie's first and second letters (indicative of words and phrases being repeated) is typical or unusual amongst different authors (perhaps looking at female/male authors, or authors from different socioeconomic backgrounds). Are, for example, some authors more formulaic than others? To what extent is the main body of the letter formulaic? Which lexicogrammatical structures are being repeated and can any trends be identified? Another line of inquiry might be to examine the high frequency of *I/you* in the Lough letters and whether this is in some way genre indicative. Would a study of other text types (narratives, diaries or spoken language) reveal similar findings? As observed by

[93] The term lexicogrammar suggests that lexis and grammar cannot be separated, but are instead two ends of the same cline.

McLelland (2007), and supported in this study, *I* is more characteristic of female authors; however, we need to learn more about how it is being employed, in which contexts, and when talking about what topics. Its use in projection clauses (as discussed in this chapter) is only part of the picture. Alternatively, the analyst might choose to investigate the high frequency of *will* and whether it is functioning in an epistemic or boulomaic sense (to show probability, or desire/volition). What other linguistic strategies are used to express modality? Are there any gender or class differences in the use of modality? Nurmi and Palander-Collin (2008), for example, found little variation in modal usage according to social differences; however, they did find some differences in usage between male and female authors, with the modals *will* and *would* being more typical of female writers. Closer investigation showed, however, that these findings varied depending on the author/recipient relationship.

For the present study, I am going to look briefly at the high frequency of mental verbs of cognition and desire, which occur after the pronoun *I* as part of a projecting clause, to see what they might reveal about the author/recipient relationships in the Lough letters.

```
ting you know all the News I think I keep you Well posted. if I did not Write but
my love to them all when you write I mail you some papers every week hope you get
for her is Kate with her in Galway I sent you some Transcripts two weeks ago when
d] we are very well at present and I thank you very much for them nice post cards
a letter from you in answer to one I wrote you the first week in September. I hope
```

Figure 2.6: Examples of non-projecting structures

From quantitative to qualitative: concordance lines

I have chosen to explore projection clauses further as the quantitative findings so far appear to suggest that these structures (or phraseological patterns) are frequently used by the Lough sisters, which may be indicative of a local grammar. The main pattern under investigation is: *I* + *V* + *you* + *(modal/aux)* + *V* (as in *I hope you will write*). I have several questions to explore: which verbs (other than *hope*) most commonly occur in projecting clauses; are there more projections of propositions (requiring a verbal response – typically expressed through a cognitive verb), or are there more projections of proposals (requiring a non-verbal response – typically expressed through a verb of desire); which auxiliary verbs most commonly follow *you* in the projected clause; does this pattern (*I* + *V* + *you* + *(modal/aux)* + *V*) attract similar text types – is it genre-indicative; is this pattern used equally by all four sisters, or does one sister use it more than the others, and finally, is this phraseology used as frequently by male and/or other female authors.

I began by carrying out a search on *I * you* ('*' is a wildcard meaning 'any word which appears in X position') in the *LOUGH Corpus*. As the findings in Table 2.15 illustrate, the search brought up <188> instances of this structure. There are three things to note at this stage. First, this search did not bring up all projection clauses, but only those where *I* occurs one word to the left of the wildcard '*'. As shown in Table 2.9, previously, *I* can sometimes occur several words to the left of the pronoun *you*, as in ***I** hope when you write again **you**…*; however, for this investigation I focused only on those (most common) structures where *I* occurs directly to the left of the mental verb. Second, the search produced only those projection clauses containing the pronouns *I/you* (separate searches would need to be carried out to identify clauses containing *I* + *you/he/she/they*, etc.). Third, not all instances of *I * you* are projection clauses. In <41> out of the <188> occurrences *you* is the Object of the main clause (rather than the Subject of a projected clause).

After having removed the non-projecting structures, there are <147> occurrences of *I * you* functioning as projection clauses in the *LOUGH Corpus*.

The auxiliary modals that most frequently follow *you* are listed in Table 2.16. The data shows that *will* is by far the most common modal used in this structure.

The verbs in Table 2.15 can be categorised in terms of the experience they are construing. For example, *assure*, *tell*, *thank* and *told* could be described as communicating or saying verbs; *dream*, *hope*, *know*, *like*, *see*, *suppose*, *think*, *want*, *wish* and *wonder* could be described as mental verbs of cognition, perception or desire; and *keep*, *mail*, *receive*, *send*, *sent*, *write*, *wrote* could be categorised as verbs of action. The data shows that the pattern *I* + *Verb* + *You* seems to attract more mental verbs, with *hope*, *know/no*, *suppose* and *wish* being the most common.

Of the <147> occurrences of *I * you* functioning as a projection clause, the most common verb to occur in this pattern is *hope*. As shown in Figure 2.5 earlier, over half of all instances of *I hope you* (<58> out of <95>) are followed by *will*. In these instances the author is placing a mild obligation on the recipient to do something – usually write, or forgive for lack of communication. Of the remaining occurrences of *I hope you*, most are standard, formulaic phrases which one might expect in any letter (*I hope you are well*, *I hope you get good health*, *I hope you can read my writing*).

These, very formulaic, projection structures are commonly found in the openings and closings of letters (as also noted by Dossena (2007) and are described by Scott and Tribble (2006) as channel maintainers, helping to sustain the lines of communication between author and recipient.

I * you	Freq.
assure	4
dreamed	1
hope	95
keep	1
knew	1
know / no	11
like	1
mail	1
received	4
see	2
send	1
sent	5
suppose	28
tell	1
thank	3
think	7
told	1
want	5
wish	11
wonder	1
write	1
wrote	3
TOTAL	**188**

Table 2.15: Search results for *I * you* in the *LOUGH Corpus*

	Modal V.	Raw freq.
I * You	can	3
	could	4
	must	1
	ought	1
	will	42

Table 2.16: Auxiliary modals following *I * you* in the *LOUGH Corpus*

The Figures below show sample concordance lines for the other main projection clauses.

```
pe you are getting along good and I Know you are trying to do the best you can I
pe you are very well yourself and I Know you are trying to do the best you can I
every as I would wish for you and I know you  are doing the best you can in your
must try and keep well if you can I know you never can stop thinking of Dear Annie
haps sooner than you think [sic]  I know you would grow young again [Page Four]
not get this before Christmas and I know you would  not be happy if you  did not
I did not write till the last for I knew you would worry and I was sure you would
```

Figure 2.7: Sample Concordance Lines for *I know you*

```
y glad I made the change although I think you and mother did not like it by your
r she is going to write this week I think you are growing smarter all the time to
er in your own dear hand writing. I think you done just splendid It was a very nice
r if those things of mine fit her I think you have been more than generous to give
I get her next letter [--damaged] I think you two ought to be very comfortable
y cold weather Julie wrote to you I think you  will have hers first I suppose she
to see such style when I go home  I think it is nonsense I think you ought to burn
```

Figure 2.8: Sample Concordance Lines for *I think you*

```
ve some very hansom under clothes I wish you could see them evry stitch of clothes
ible cold and talk of snow drifts I wish you could see some of them this last week
und the skirt 20 yds all together I wish you could see it I think I will send you a
ie will come to see you often and I wish you could go see her sometimes give my
icture in my next letter if I can I Wish you would try an have some of your picturs
letter and glad you are all well  I wish you would write oftener but I suppose you
ve but I take pleasure in sewing. I wish you was near so I could help you  you
```

Figure 2.9: Sample Concordance Lines for *I wish you*

```
would wish so much to see you all I suppose you all felt bad for Parnell it was too
s 16 years she has only two girls I suppose you all read about our presidents death
mas  how is the winter over there I suppose you are bussey getting ready for xmis
ve devotion evenings during Lent. I suppose you are always busy. how is Maggie does
ends here are very well and Alice I suppose you are going to school and is at home
Mary gets good health Dear Mother I suppose you were worried some about that letter
that ye well spend a happy xmas   I suppose you will be getting good xmas presents
he station to see me off now Mary I suppose you will be tired reading all this so I
hers I would do the same With his I suppose you will think I am not goeing to say
lines hoping to find you all Well I suppose you will think that I am never goeing
ear all about the old  neighbors. I suppose you will never get over been lonesome
there some time yourself yet but  I suppose you would hate to give up the old place
wish you would write oftener but  I suppose you do quite a  lot of writing to the
November and sent you some papers I suppose you got them all right I am sending you
e snow drifts is not all gone yet I suppose you have all the planting done at home
would blow it away some time but  I suppose you Keep it repaired now and again I
g well considering the hard times I Suppose you must have heard of the hard times
lease  give her my loving regards I suppose you people over there do not fast in
all power to do as he thinks best I suppose you reed lots about this country I hope
planted  no potatoes we buy them. I suppose you still do the same with yours I was
```

Figure 2.10: Sample Concordance Lines for *I suppose you*

Looking at the concordance lines for *I know you* and *I think you*, first of all, it appears that *know* and *think* in these clauses are being used as subjective modality markers, rather than true mental projection verbs. These phrases seem to be used when expressing sympathy, or as a way of showing solidarity. The author, in these lines, places themselves in the position of the recipient, imagining their behaviour, what they are doing and how they are feeling. In the case of *I wish* (specifically, *I wish you could see* <freq. 3>) this empathy is reversed and the recipient is invited to imagine something from the author's perspective. Other instances of *I wish* are used to admonish – *I wish you would write oftener* and *I wish you would try to have your photo taken*.

Whereas *wish* is being used to express boulomaic modality (the author, in these instances, is expressing a desire for the recipient to do something (*write oftener*) or experience something (*see her*)), *suppose*, on the other hand, is being used to express epistemic modality. With a degree of certainty, albeit hedged, the author is predicting what the recipient is thinking, feeling or doing. The use of epistemic modality, in these occurrences, emphasises, strengthens and reinforces familial bonds – bonds that are based on past, shared experiences between the two participants. In saying *I suppose you were worried some about that letter* the author is doing more than empathising – they are showing a connection with the recipient which is based on previous and existing knowledge between the two correspondents, which transcends space and time – the message being: 'based on past experiences, and knowing you in the intimate way that I do, my guess is that you are feeling worried'.

The type of projection taking place in these concordance lines (except for instances of *wish*) is a proposition, where the mental verb is one of cognition (*know*, *think*, *suppose*). These projections of propositions require a verbal response, placing a mild obligation on the recipient to (verbally) acknowledge and address the points being raised. These clauses, then, help to facilitate the interactive nature of the letter – establishing and maintaining a dialogue between the two participants. However, as discussed previously, the most frequently occurring verb in the pattern *I * you* is *hope*, often used to project a proposal (i.e. something which requires a non-verbal response, as in *I hope you will try and be very happy and enjoy yourself*). A closer look at the distributional trends of these *I * you* structures shows that whereas *I hope you* more typically appears in the openings and closing of the letters, *I think/know/suppose you* tends to occur more frequently in the main body.

Having carried out a search on the projection clause *I * you* in the *LOUGH Corpus* I then carried out the same search, but this time looking at each subcorpus to see whether one sister uses this structure more than others. The same search was also carried out using the *MALE Ref.* and *FEMALE Ref.* corpora to see whether any gender differences (concerning the use of projection clauses) could be identified. The findings are shown in Table 2.17. Looking at the normalised figures, the data suggests that there is no significant difference in the usage of this structure between Annie, Julia and Alice, although Lizzie seems to use *I* you* much less than her siblings. The data also suggests that female authors use this structure more than male authors; however this is a very general and tentative finding as both reference corpora contain a mixture of authors from different socioeconomic backgrounds, making it difficult to draw any specific conclusions. Indeed, the same search, but this time using a much larger (450 million word), contemporary reference corpus (the *Bank of English*), showed that this structure most commonly occurs in spoken language (see Table 2.18 – 'brspok' refers to the British spoken language subsection of the corpus and 'usspok' refers to the US spoken language subsection), which could mean that the differences in usage of *I * you* are more indicative of differences in educational background, with letters that adopt a more colloquial, speech-like style making greater use of projection clauses.

	Freq. of I * You	Normalised
NLC	102	5,39
JLC	57	4,65
ELC	10	2,87
ALC	17	4,74
LOUGH Corpus	186	**4,86**
FEMALE Ref.	35	**2,12**
MALE Ref.	23	**1,11**

Table 2.17: The pattern *I * you* within the *LOUGH Corpus* and reference corpora

Corpus	**I+1,2you+will**	**I+1,2you+can**	**I+1,2you+could**
usephem	35,4	4	2,6
brephem	16,6	2,4	1,1
usspok	13,3	58,3	18,8
brbooks	7,4	7,5	4,8
brspok	4,6	53,5	27,7
usbooks	4,5	6,4	3,7
sunnow	3,8	4,1	2,9
strathy	3,1	3,8	2,6
brmags	2,4	4,7	2,2
indy	1,9	2,7	1,2
npr	1,7	11,3	6
usacad	1,6	1,6	0
guard	1,5	2,7	1,4
times	1,5	2,8	1,4
oznews	1,3	2,2	2,2
newsci	1	1,1	0,5
bbc	17	1,6	0,6
usnews	9	1	0,8
wbe	5	0,8	0,5
econ	5	0,3	0,1

Table 2.18: The pattern *I * you* within the *Bank of English*

Discussion and conclusions

At the beginning of this chapter I proposed a method of inquiry based on the theory and techniques of corpus linguistics. Taking simple frequency data as the starting point, I was alerted to certain linguistic patterns, which an ordinary reading of the letters may not have allowed. The language contained within the letters was first taken out of its context; it was reorganised to reveal recurring linguistic features worth further, more qualitative, investigation. The findings were then considered within the situational and cultural context of international migration to try and build a picture of how, through letters, family bonds were changed and maintained over space and time.

The approach this study adopts starts with individual words and then examines how those words behave in sentences. What emerges is a specific phraseological pattern (*I* + *V* + *you* + *(modal/aux)* + *V*), which, further comparative investigations seem to suggest, is used more by female authors than male authors. These projecting structures place the recipient (*you*) – in this case, usually the mother or sister, Mary – as the Subject of the projected clause. However, at the same time, they also place the author (or more specifically the author's expectations, needs or desires) in the sentence initial position. In other words, these structures lead with some expectation of the author that is highlighted before we reach the main point of the sentence, which requires action, whether verbal or non-verbal, on the part of the recipient. The function of these clauses is to project an imagined narrative onto those back home, arguably serving to maintain a psychological link between the emigrant and their family in Ireland. It is through these, somewhat mundane, repeated phraseological patterns that familial relationships are strengthened and reinforced. The frequency of projection clauses in the Lough letters certainly warrants further investigation and in chapter three I will use corpus query language (CQL) to extract all instances of these structures (including, for instance, those which contain the pronouns *he*, *she*, *they*, *we*, *it*) to examine, in more detail, their different functions.

The approach taken in this chapter is very selective. As mentioned earlier, the initial quantitative investigations highlighted several possible lines of inquiry; however, I chose to follow just one of those, whilst ignoring others. What this approach does offer, however, is a clear, data-led rationale for choosing to examine certain linguistic features in the first place. The numbers themselves are not problematic, nor, necessarily, are the statistical measures or tests that are applied. What, arguably, is problematic are the research questions that are asked in the first place, the data that is used to explore those questions, and/or the conclusions that are later inferred from the results. McLelland's (2007) study, discussed earlier, shows how statistics cannot be taken at face value, but should be tested, re-tested and tested again in different ways, against different data sets and by scholars from different disciplinary perspectives. Each line of inquiry will provide different findings, but combined will allow for a fuller, more complete profiling of the female experience of migration. The present study found that a certain pattern appears to be used more by female authors; however until this finding is tested against other data sets (taking into account factors such as social class, educational background, frequency of writing and so on) it is difficult to speak conclusively about the results. (In chapters five and six I will propose a method of markup for capturing extra-linguistic information, which will allow users to test hypotheses whilst taking into account sociobiographic variables.) Nevertheless, the methodology proposed here is transparent and replicable. The results can be tested, challenged, rejected or confirmed and it is through this process that nuances relating to gender history can begin to emerge.

In many ways this chapter has brought up more questions than it has provided answers, but one of its main aims was to demonstrate how quantitative methods of analysis might tease out interesting linguistics features for further (quantitative and qualitative) analysis. This chapter has put forward a complementary methodology for examining gender history. It has highlighted some of the possibilities and challenges of using quantitative methods to support, build-on or challenge more qualitative research. Equally, however, it is hoped that some of the quantitative findings discussed here will be taken up by scholars using more qualitative approaches, providing new layers of meaning to the quantitative findings and giving new insights into the individual emigrants who the numbers represent.

CHAPTER THREE

Case study: A closer look at projection structures in the *LOUGH Corpus* using Corpus Query Language

Introduction

In this case study I will look more closely at the use of the pattern *Pronoun + Verb + Pronoun* (as in *I hope you*, *you think I*, *she knows he*) in the *LOUGH Corpus*, which, in chapter two, was identified as being a statistically significant feature of the Lough letters.[94] I will investigate how these clauses – described as projection structures – function and how they contribute to the interactive nature of letters, helping to strengthen and maintain familial bonds over time and distance.

In carrying out this analysis I draw on the concept of intersubjectivity in language. For Traugott and Dasher, intersubjective meaning comes directly from the interaction between a speaker/writer (SP/W) and an addressee/reader (AD/R), and can be characterised as the 'SP/W's attention to [and awareness of] the AD/R as a participant in the speech event' (2002, p. 22). Intersubjective meaning encodes the SP/W's point of view whilst at the same time discursively positioning the AD/R, assigning them a role to play in the 'unfolding of the discourse' (Thompson 2012, p. 80). Similar to Thompson (2012), who draws on the work of Bakhtin (1986), this study takes a broad, discoursal approach to intersubjectivity, viewing all discourse as dialogistic – that is, 'constructed

[94] Note that the post-verbal pronoun is Subject case so as to exclude instances such as *I know her* or *I see them*.

fundamentally in terms of exchanges between interactants in communicative events in which each interactant shapes their message to accommodate and affect the other' (Thompson 2012, p. 78). There are, of course, interpretative limitations when working with one-directional correspondence collections, such as the Lough letters. For one, it is simply not possible to know how the recipient of the letter responded to its content. However, an analysis of the the linguistic choices found within the letters will 'reflect the writer's [or author's] expectations about what the addressee [or recipient] may bring to the text and the kinds of response that the text will elicit from the addressee' (Thompson 2012, p. 80).

One of the ways that intersubjectivity is realised in language is through what Halliday and Matthiessen (2004) describe as projection structures (e.g., ***I know you*** *are doing the best you can* or ***I suppose you*** *did not mind* (*JLC*, 11 May n.d.)). In the letters, these structures often explicitly speak to the recipient of the letter – *you* – and have the ability to project the author's expectations, desires, or beliefs onto the recipient, thus helping to construct what Thompson and Thetala describe as 'reader-in-the-text' (1995, p. 103). These structures not only express the author's point of view, but also construct a recipient (or reader-in-the-text) 'with certain attitudes, knowledge, assumptions, status, etc.' (Thompson 2012, p. 80). They anticipate reactions and seek to elicit certain responses, thus contributing to the interactive nature of the letters and helping to strengthen the relationships those letters embody.

This chapter will examine intersubjectivity – as realised through the use of projection structures – in the Lough letters. It will use computational and corpus methods to, first, identify and extract these linguistic structures before giving a quantitative overview of how they are being employed by the Lough sisters. A closer, more qualitative analysis will then examine the communicative function of these projection structures and how they contribute to maintaining a psychological link between author and recipient.

Previous studies

Previous research has explored the use of pronouns and evidential verbs in personal correspondence to see what these linguistic features might reveal about the author/recipient relationship (evidential verbs are those which express the writer's 'attitude to knowledge', such as *know*, *think*, *believe*) (see Chafe 1986, p. 262 cited in Palander-Collin 1999, p. 124). Sairio (2005), for instance, explores levels of linguistic involvement in letters by Samuel Johnson. Drawing on the work of Chafe (1985), Sairio makes a distinction between ego- or self-involvement (typically realised through first person pronouns); interpersonal-involvement between author and recipient (typically realised through second person pronouns); and the author's involvement with the topics being discussed in the letter, for which a range of linguistic devices might be employed (see, for example, Simpson's 1993 work on style and point of view). Focusing primarily on ego- and interpersonal-involvement, Sairio's findings show how the use of first and second person pronouns as well as evidential verbs are 'a relevant indicator of the closeness of the relationship' (2005, p. 33): the closer the relationship the more likely it is that these linguistic devices will be used. Looking at letters by Samuel Johnson to two of his correspondents – Mrs Thrale (a close friend) and Lucy Porter (Johnson's step-daughter) – Sairio found that the level of linguistic involvement generally decreased over time, with fewer evidential verbs found in later letters to both Thrale and Porter. Additionally, Sairio's study found a decrease over time in first person pronouns (indicating ego-involvement) in letters to Mrs Thrale and a decrease in second person pronouns (indicating interpersonal involvement) in letters to Lucy Porter. This general decline in levels of involvement may point to possible changes in the relationship between the correspondents, although, as Sairio points out, other factors could also be responsible for the change (e.g., Johnson's age or life situation) (2005, p. 32).

Also examining the use of first person evidential phrases (such as *I think*), but this time focusing on seventeenth-century letters from the *Corpus of Early English Correspondence* (CEEC),[95] Palander-Collin (1999, p. 139) found that 'women's personal letters show a more involved style than men's letters', with female authors using significantly more first person evidential verbs than their male counterparts (see also studies by Nurmi and Palander-Collin (2008), and Säily, Nevalainen and Siirtola (2001) both of which suggest gender-based variation in the use of pronouns). A more recent study by Palander-Collin (2009), which investigates sixteenth century correspondence from CEEC (specifically letters by Nathaniel Bacon, a younger son of Sir Nicholas Bacon), found that 'the frequency of self-mention and addressee inclusion varies according to the addressee' with first and second person pronouns occurring '…more often when writing to social inferiors, equals and family members, and less often to social superiors' (p. 65).

[95] *The Corpus of Early English Correspondence* (CEEC) (1993-present) was compiled by Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A. and Palander-Collin, M. in the Research Unit for Variation, Contacts and Change in English (VARIENG) at the University of Helsinki. More information about the corpus can be found on the project website: http://www.helsinki.fi/varieng/domains/CEEC.html [Accessed 1 May 2015].

The studies outlined here show that the level of linguistic involvement in personal letters (as realised through the use of first and second person pronouns and evidential verbs) varies depending on factors such as the author/recipient relationship; the gender of the author and/or recipient; and, quite possibly (although more research is needed here), the amount of time that has passed (with earlier correspondence within a letter series tending to show greater involvement than later correspondence).[96] All of these are potentially interesting areas to explore further with reference to emigrant letters. This is especially the case with respect to the last point, given the immense pressure emigrants and their loved ones were under to maintain family relationships across distance and time.

Building on this previous research, the present study not only investigates the frequency with which the Lough sisters use linguistic indicators of involvement, but it also seeks to explain the function of those linguistic features and how they helped to construct and maintain the relationships embodied within the Lough letters. To do this, the use of pronouns and evidential verbs will be examined within their wider phraseological context, focusing specifically on their use within what Halliday and Matthiessen describe as types of projection.

Types of projection

In the previous chapter, the pattern *I + V + you + Md/Aux + V* (as in, *I hope you will write* or *I suppose you will never get over been* [*sic passim*] *lonesome* (*NLC*, n.d.)) was found to be particularly frequent in the Lough letters. Within systemic functional grammar (e.g., Halliday and Matthiessen 2004) this pattern is

[96] Unfortunately there are no similar studies to report concerning letters between working-class-equivalent and very moderately educated people, who are also kin, siblings or child-parent relationships. Therefore, it is necessary to be tentative when discussing relevant norms (relevant to the Lough letters) from Johnson and other elevated figures.

described as a type of projection. Other studies have described these structures as clausal epistemic parentheticals (see Huddleston and Pullum 2002; López-Couso and Méndez-Naya 2010 & 2011; and Thompson and Mulac 1991); comment clauses (see Quirk et. al. 1895; and Brinton 2008); or metadiscursive phrases (Ädel , 2012). Projection structures consist of two main components: the projecting clause (*I hope*) and the projected clause (*you will write*). In these structures the primary (projecting) clause (*I hope*) sets up the secondary (projected) clause (*you will write*) as the representation of the content of either what is thought, or what is said (Halliday and Matthiessen 2004, p. 377).

There are three main areas to consider when examining projection structures. The first is to do with the **level of projection**. The projection may be a representation of what is thought, as in *I think she is a good girl* (*ELC*, 7 March 1876) – and, hence, depict ‘ideas’ – or the projection may be a representation of what is said, as in *I told Annie it would pay her to move down on Main Street* (*JLC*, 2 December 1889) – and, hence, capture ‘locutions’ (Halliday and Matthiessen 2004, p. 443). The second area to consider relates to the **mode of projection**. Is the idea or locution represented as a direct quote (as in, *she said, ‘I am expecting a letter’*) or as a report (as in, *she said she is expecting a letter*)? Whereas quotations can stand independently of the projecting clause, reports are dependent on the projecting clause and cannot stand on their own. The third area – which is most relevant to this chapter – is the **speech function of the projection**. Halliday and Matthiessen make a distinction between the projection of propositions and the projection of proposals as follows:

> [P]ropositions, which are exchanges of information [typically statements or questions], are projected mentally by processes of cognition – thinking, knowing, understanding, wondering, etc. ... [P]roposals, which are exchanges of goods-&-services [typically offers or commands], are projected mentally by processes of desire (2004, p. 461).

Both propositions and proposals have different response-expecting speech functions. Propositions generally require a verbal response from the recipient: for example, the recipient of *I Know you never can stop thinking of Dear Annie* (*NLC*, 18 October 1928) may agree or disagree with this statement. Proposals generally require a non-verbal response from the recipient: so in the example *I hope you will try and take very good care of yourself* (*NLC*, 14 July 1918), the recipient may choose to follow up on this (albeit indirect) command and eat/rest, or not. In the case of proposals, then, what is effectively a command – *take care of yourself* or *keep the children in school* – can be expressed as a statement *I hope you will try and take very good care of yourself* (*NLC*, 14 July 1918) or *I hope you keep them to school all you can* [*sic passim*] (*NCL*, 10 December 1902). Through presenting a command, usually an imperative, as a statement, usually a declarative (a process which is described by Halliday and Matthiessen as mood metaphor), the speaker/writer is able to personalise the command by incorporating a Subject and a Finite, thereby opening up the possibility for negotiation and interaction (for more information about the use of mood metaphor in correspondence see Wei-Ling Wee (2009)).

An analysis of projection structures will, therefore, reveal something about intersubjective meaning: that is, how the author interacts with their intended recipient and the type of response they expect – whether that is a verbal response requiring the recipient to agree, empathise or object etc., or a non-verbal response requiring the recipient to carry out an action of some description. Both types of interaction (the projection of propositions and the projection of proposals) involve the recipient in different ways, potentially revealing something about the author/recipient relationship.

This chapter will investigate the use of these projection structures in the Lough letters. It first uses Corpus Query Language (CQL) to identify the structures and then uses corpus tools to capture the information used by the author to create dialogue between the author and recipient (proposition), or to negotiate a desired action (proposal). Such information includes: who or what is in the position of Subject in the projecting clause; what is being projected (ideas/thoughts or locutions/speech), and what is the speech function of the projection. I also explore whether there is a correlation between the type of projection used and the author/recipient relationship.

Methods

As in chapter two, to prepare the letters for corpus analysis, they first had to be digitised and then saved in plain text format. The corpora were then loaded into Sketch Engine,[97] which automatically assigns each word a Part of Speech (POS)

[97] Kilgarriff, A. and Kosem, I. (2012) Corpus Tools for Lexicographers. In S. Granger and M. Paquot (eds.), *Electronic Lexicography*. New York: Oxford University Press. pp. 31-56. Available from: http://www.sketchengine.co.uk.

tag using the Penn Treebank tagset. This allowed me to specify the parts of speech I wanted to search for:

- [tag="PP|PP$|NP"] to select all personal pronouns (PP) and/or possessive pronouns (PP$) and/or proper nouns (NP). (The '|' symbol means 'and/or'.)
- [tag="V.."] to select all forms of a verb.
- [tag="RB"] to select adverbs.
- [tag="MD|V.."] to select modal verbs (MD) and/or all forms of a verb (V…).
- [word="XXX"] to select a specific word (where 'XXX' is substituting the word in question).
- [] to select any word which appears in X position.

Using Corpus Query Language (CQL) it was then possible to create search queries that allowed me to extract the types of projecting structures described earlier. Six main patterns were investigated:

1. [tag="PP|PP$|NP"]

I began with a search for all personal and possessive pronouns (*I*, *you*, *he*, *she*, *me*, *his*, *her* etc.)[98] and/or all singular proper nouns (*John*, *Mary*, *Maggie*, *Annie* etc.), to see how often the Lough sisters refer to themselves and others in the letters. Initially, I included plural proper nouns in this search, hoping to identify references to families – *the Deevys* or the *O'Hanlons*, for instance. However, it soon became apparent that, due to a lack of punctuation in the Lough letters

[98] One of the reasons why possessive pronouns were included in the search queries was to capture any instances where the projected clause contains a determiner (possessive pronoun) + noun (as in, I hope *his brother* is well). As it happens, possessive pronouns are not used in this way in the Lough letters.

(there is just one apostrophe in the entire corpus and very few full stops), this search produced mostly possessive structures (*Alices*, *Annies*, *Gods*). Plural proper nouns were therefore not included in this search.

2. [tag="PP|PP$|NP"] [tag="V.."]

The second search query identified all pronouns and/or proper nouns followed by any verb form to see which verbs tend to co-occur with which Subjects, thus revealing something about who is thinking, feeling, seeing or doing. I decided to search for any verb form (rather than specifying tense and aspect) as I wanted to keep the search criteria as open and inclusive as possible, so as not to miss potential syntactic variations such as pronoun followed by past participle (as in, *I done* or *Maggie seen*) – a structure that occurs (albeit infrequently) in the Lough letters. It was important – given that the letters used in this study are written by lower-class, minimally educated authors – that the search criteria were as flexible as possible. In describing the authors as 'minimally educated' I am drawing on the work of Fairman (2009; 2012) who argues that certain linguistic features – chaining and a lack of embedding, lack of punctuation, and more anglo-saxon than latinate words (all of which are found in the Lough letters) – might suggest what he describes as mechanical, or minimal, schooling (Fairman 2009; 2012).

3. [tag="PP|PP$|NP"] [tag="V.."] [tag="PP|PP$|NP"]

This search query identified the projection structure this study is interested in: *I hope you*, *I wish you*, *I know you*, etc. However the search did not identify projecting clauses containing adverb/verb combinations (as in, *I always hoped*), nor did it identify negative structures (as in, *I do not think*). Therefore, additional searches (see 4, 5 and 6 below) were carried out to identify and extract these patterns.

4. [tag="PP|PP$|NP"] [tag="RB"] [tag="V.."] [tag="PP|PP$|NP"]

This search identified projection structures containing adverbs, as in *I really thought Mag had more sense than that* or *I often wish you had some nice little place to live*.

5. [tag="PP|PP$|NP"] [tag="MD|V.."] [word="not"] [] [tag="PP|PP$|NP"]

This search identified all projection structures containing a *modal/auxiliary verb + not*, as in *she ought not tell you* or *I do not think he*. However, it did not account for those instances where negation is expressed through a contracted form (see 6 below).

6. [tag="PP|PP$|NP"]
[word="dont|dident|didnt|doesnt|cant|couldnt|wouldnt|wount|wont|isnt"] []
[tag="PP|PP$|NP"]

As previously mentioned, punctuation rarely occurs in the Lough letters. It is a similar case in the two reference corpora (detailed in chapter two), with just six apostrophes in the *MALE Ref. Corpus* and none in the *FEMALE Ref. Corpus*. A search for apostrophes, therefore, would not necessarily produce instances of contracted forms. Not only that, spelling variations amongst the different authors meant that a search for *didnt* would miss instances of *dident* and a search for *wont* would miss instances of *wount*. It was therefore necessary to first identify which contracted forms are used to express negation in the *LOUGH Corpus* and the two reference corpora. This involved examining the wordlists for each corpus. *Shouldnt*, *mustnt* and *arent* do not occur in any of the three corpora. Contracted negative structures that do occur are: *dont*, *dident*, *didnt*, *doesnt*, *cant*, *couldnt*, *wouldnt*, *wount*, *wont*, *isnt*. These contracted forms were thus incorporated into the search query.

What follows is a summary of the key findings. In most of the tables there is a column entitled 'Freq.' which provides the raw (or actual) frequencies and a column entitled 'Norm.' which provides the normalised frequencies. Normalised frequencies allow meaningful comparisons to be made across datasets of different sizes. It is calculated by dividing the raw frequency by the total number of words in the corpus, times 1000, giving an average frequency of a particular word or phrase per 1000 words.

Findings

In Table 3.1 the first column, 'CQL Ref.', corresponds to the six CQL search queries outlined in the previous section, with the second column showing the lexicogrammatical patterns that each search extracts. CQL Ref. 1, for example, uses the CQL search query [tag="PP|PP$|NP"] to extract instances of all pronouns and/or proper nouns (represented as *Pr/N* in column two). The first section of the table gives the raw and normalised frequencies for the *LOUGH Corpus* as a whole as well as the *MALE Ref.* and *FEMALE Ref.* corpora. The second section of the table (directly underneath) gives the raw and normalised frequencies for each of the Lough sisters.[99] Throughout this chapter I will be referring to the normalised frequencies, unless otherwise stated.

Looking at CQL Ref. 1, first of all, the normalised frequencies suggest that the *LOUGH Corpus* contains more pronouns and proper nouns (*I*, *you*, *he*, *she*, *Maggie* etc.) than both the *MALE* and *FEMALE* reference corpora: <203.65> occurrences in the Lough letters versus <155.55> for male authors and <189.93> for female authors. There appears to be a general tendency for female authors to use pronouns/proper nouns more frequently than their male counterparts, which would support previous studies that have shown similar gender differences in pronoun usage.[100] Moving on to CQL Ref. 3, *Pr/N + V + Pr/N* (as in, *I wish you*), the findings indicate that this pattern is used significantly more by the Lough sisters, <18.41>, when compared with both the *MALE* and *FEMALE* reference corpora <6.51> and <10.65> respectively. Again, this pattern appears to be gender specific, with female authors using the structure almost twice as frequently as their male counterparts and, in the case of the Lough sisters, almost three times more than the male authors. The first (and older) sisters to emigrate (Lizzie and Alice) use this structure the least, <17.47> and <15.97> respectively. The last (and younger) sisters to emigrate (Annie and Julia) – who also happen to write most frequently – use it the most, <20.69> and <23.90> respectively.

[99] The accumulative figures for the Lough sisters will not necessarily correspond with the total figures for the *LOUGH Corpus*. This is because the *LOUGH Corpus* includes letters where the author has not yet been established and, as such, these letters cannot be assigned to a particular sister.

[100] See studies by McLelland (2007), Nurmi and Palander-Collin (2008) and Säily et. al. (2011).

			LOUGH		MALE		FEMALE	
CQL Ref.	Pattern	Example	Freq.	Norm.	Freq.	Norm.	Freq.	Norm.
1	Pr/N	*I*	9072	203,65	24271	155,55	6793	189,93
2	Pr/N + V	*I wish*	3695	82,95	7885	50,53	2212	61,85
3	Pr/N + V + Pr/N	*I wish you*	820	18,41	1015	6,51	381	10,65
4	Pr/N + Adv + V + Pr/N	*I never thought I*	39	0,88	68	0,44	23	0,64
5	Pr/N + Md/Aux + not + V + Pr/N	*she did not think he*	50	1,12	109	0,70	27	0,75
6	Pr/N + (Md/Aux + not) + V + Pr/N	*I dont think I*	23	0,52	26	0,17	10	0,28

			ELC		ALC		NLC		JLC	
CQL Ref.	Pattern	Example	Freq.	Norm.	Freq.	Norm.	Freq.	Norm.	Freq.	Norm.
1	Pr/N	*I*	692	236,99	746	212,72	4134	212,82	2841	255,21
2	Pr/N + V	*I wish*	274	93,84	308	87,82	1634	84,12	1219	109,50
3	Pr/N + V + Pr/N	*I wish you*	51	17,47	56	15,97	402	20,69	266	23,90
4	Pr/N + Adv + V + Pr/N	*I never thought I*	6	2,05	3	0,86	17	0,88	12	1,08
5	Pr/N + Md/Aux + not + V + Pr/N	*she did not think he*	5	1,71	1	0,29	21	1,08	21	1,89
6	Pr/N + (Md/Aux + not) + V + Pr/N	*I dont think I*	3	1,03	5	1,43	8	0,41	4	0,36

Table 3.1: Frequencies for the CQL searches

Looking at CQL Ref. 4, *Pr/N + Adv + V + Pr/N* (e.g., *I never thought I*), the data suggests that the Lough sisters, and female authors more generally, tend to make greater use of adverbs within this pattern when compared with male authors: <0.88> and <0.64> for the *LOUGH Corpus* and the *FEMALE Ref. Corpus* versus <0.44> for the *MALE Ref. Corpus* – the main adverbs being *always*, *often*, *never*, *ever*. Although the difference in frequencies between the *MALE* and *FEMALE* reference corpora is not hugely significant, the Lough sisters nonetheless use adverbs exactly twice as often as the male authors. This is mainly due to Lizzie who makes particular use of adverbs in her writing, with a normalised frequency of <2.05> compared with <0.86> for Alice, <0.88> for Lizzie and <1.08> for Julia. There is a similar trend with regard to the use of negation. The *LOUGH Corpus* has twice as many patterns containing negation than the *MALE Ref. Corpus* – <1.64> and <0.87> respectively. Overall, female authors appear to use negation in these patterns slightly more than their male counterparts – a normalised frequency of <1.03> in the *FEMALE Ref. Corpus* – although, again, the difference is not especially significant.

Having established a general overview of the frequency and distribution of these patterns, the next step was to establish exactly how many of these search outputs were, in fact, projection structures. The search query CQL Ref. 3 (*Pr/N + V + Pr/N*), for example, brought up instances such as *I thank you* and *you sent me* (as in, *I thank you for the papers you sent me*), *you gave her* (as in, *you gave her a nice name*), and *I let her* and *her read your* (as in, *I let her read your last letter*), all of which are not functioning as projections, but are instead straightforward Subject/Verb/Object constructions. It was necessary, therefore, to sift through each search output qualitatively, identifying those structures that were projecting and those that were not. At this point in the study, any instances that were not functioning as projection structures were discounted. Table 3.2 summarises the results. CQL Ref. 3 gives the raw frequencies for the projection structure *Pr/N + V + Pr/N* (as in, *I hope you…*), CQL Ref. 4 gives the raw frequencies for those projection structures containing adverbs (as in *I often wonderd* [*sic passim*] *she…*) and CQL Refs. 5 and 6 give the raw frequencies for projection structures containing negation (as in *you need not say you forget* and *I dont think she ever will*). The 'TOTAL' column provides the total raw frequency of projection structures for each sister and the 'Norm.' column provides the normalised figures.

	CQL Ref. 3	CQL Ref. 4	CQL Ref. 5	CQL Ref. 6	TOTAL	Norm.
ELC	23	4	2	1	30	10.27
ALC	35	2	1	4	42	11.98
NLC	286	7	3	5	301	14.71
JLC	199	6	1	4	210	15.06

Table 3.2: Frequencies for projection structures by sister

Looking at the 'Norm.' column, the data show that Annie and Julia – the younger of the Lough sisters, who are the last to emigrate, but are the most frequent writers – make greater use of projection structures in their letters. Lizzie and Alice – the two older sisters, who are the first to emigrate, but rarely write home – make less use of this structure in their letters. The correlation between frequency of writing and the use of projection structures might suggest that this pattern is genre specific and/or indicative of a more experienced writer, which in turn could suggest differences in educational background between the four sisters. These are, however, very tentative hypotheses at this stage.

Having identified and extracted the patterns that realise the projection structures this case study is interested in, the next step was to take a closer look at the lexis. I started by examining the use of pronouns and proper nouns. Table 3.3 provides normalised frequencies for pronouns/proper nouns in the position of Subject in both the projecting clause (the *I* in *I hope*, described in the table as 'P-ing') and the projected clause (the *you* in *you will write*, described in the table as 'P-ed'). The most frequent pronouns are underlined and in bold.[101] 'PN' refers to instances of proper nouns.

	ELC		ALC		NLC		JLC	
	P-ing	P-ed	P-ing	P-ed	P-ing	P-ed	P-ing	P-ed
I	**7.88**	1.37	**10.27**	1.14	**12.80**	0.78	**11.26**	2.30
you	1.37	1.37	0.29	**5.13**	1.08	**6.99**	1.43	**6.17**
he	0.00	0.00	0.00	0.86	0.00	0.44	0.29	0.57
she	0.68	**3.77**	0.29	0.86	0.34	1.56	1.08	2.58
they	0.00	0.00	0.86	1.14	0.10	0.93	0.14	0.36
we	0.00	0.34	0.00	0.57	0.05	0.29	0.07	0.29
PN	0.34	1.03	0.29	1.14	0.34	2.49	0.79	1.29
it	0.00	1.03	0.00	0.29	0.00	0.64	0.00	0.65
me	0.00	0.68	0.00	0.29	0.00	0.39	0.00	0.65
her	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00
him	0.00	0.34	0.00	0.00	0.00	0.00	0.00	0.00
them	0.00	0.00	0.00	0.57	0.00	0.15	0.00	0.00
us	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00

Table 3.3: Frequencies for pronoun usage in projection structures

As one might expect, *I* is most typically the Subject of the projecting clause across all four sub-corpora (*ELC*, *ALC*, *NLC* and *JLC*). Lizzie and Julia (the oldest and youngest of the sisters to emigrate) also show a relatively high frequency of *you* in this position (as in, *you know I am thinking of you* (*JLC*, n.d.

[101] The pronouns *me*, *her*, *him*, *them*, *us* are less common, but they are part of projection structures nonetheless (as in, *I Knew I would read of Mr Fitzs death because you told me was very feeble* (*NCL*, 14 August 1919)).

December 1890) and *you would not think you eaver* [*sic passim*] *seen her* (*ELC*, 7 March 1876)). This places the recipient of the letter in sentence-initial position, often making their (the recipient's) thoughts, needs, wants or desires the central theme. For Alice, Annie and Julia, the most common pronoun in the position of Subject of the projected clause is *you*, whereas, for Lizzie, *she* is most frequently the Subject of the projected clause. This may reflect Lizzie's role within the family as the older sister. In these structures Lizzie projects her thoughts and desires onto her younger siblings (*I thought she Would look like me but I gess* [*sic passim*] *she Wount* (*ELC*, 7 March 1876) and *I think she ought to be home* (*ELC*, 7 March 1876)) – thereby adopting and asserting the caring, authoritative, older sister role. Figure 3.1 summarises the interactants or 'readers/writers-in-the-text' (Thompson 2012, p. 83) involved in these structures.

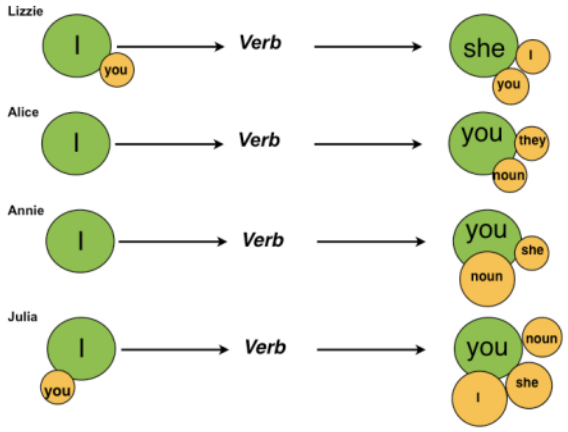


Figure 3.1: The main (pro)nouns found within projection structures

The next stage was to identify which verbs are used in these structures so as to ascertain what is probably being projected – an idea/thought or a locution/speech.[102] Additionally, if the projection is an idea I wanted to see whether it was a proposition (i.e. an exchange of information, thus creating

[102] It should be noted, however, that speech is often reported by 'thought' verbs. *Jon thought it was wrong* may be used to report Jon saying it was wrong, for example.

dialogue between the interactants) or a proposal (i.e. an exchange of goods and services where the aim is to negotiate a particular action or outcome). To investigate this, a qualitative study of each search output (or concordance line) was needed. Table 3.4 summarises the main findings. The 'Verb' column lists the most frequent lemmas found in these structures. The lemma includes all forms of the verb (regardless of tense and aspect), so HOPE would include *hope*, *hoped*, *hopes*, *hoping* etc. The most frequent lemmas are underlined and in bold.

Verb	ELC	ALC	NLC	JLC	TOTAL
Idea: proposal					
HOPE	1.03	**4.85**	**7.43**	**4.59**	**17.90**
WANT	0.00	0.29	0.15	0.57	1.01
WISH	0.34	**1.71**	0.69	0.72	**3.46**
LIKE	0.00	0.00	0.05	0.00	0.05
TOTAL	1.37	6.85	8.32	5.88	**22.42**
Idea: proposition					
KNOW	0.00	0.29	0.73	**1.22**	**2.24**
THINK	**4.11**	1.14	1.51	**2.65**	**9.42**
SUPPOSE	1.71	**1.71**	2.30	**1.94**	**7.66**
GUESS	0.34	0.00	0.00	0.00	0.34
HEAR	0.68	0.00	0.00	0.07	0.75
SEE	0.00	0.00	0.05	0.29	0.34
REMEMBER	0.34	0.00	0.05	0.00	0.39
WONDER	0.00	0.00	0.20	0.00	0.20
DREAM	0.00	0.00	0.00	0.07	0.07
REALISE	0.00	0.00	0.00	0.07	0.07
EXPECT	0.00	0.29	0.00	0.00	0.29
CONSIDER	0.00	0.00	0.00	0.14	0.14
FANCY	0.00	0.00	0.00	0.07	0.07
TOTAL	7.18	3.43	4.84	6.52	**21.98**
Locution					
SAY	0.00	1.14	0.68	1.22	3.04
ASK	0.68	0.00	0.10	0.14	0.92
TELL	1.03	0.57	0.78	0.50	2.88
INFORM	0.00	0.00	0.00	0.14	0.14
ASSURE	0.00	0.00	0.00	0.43	0.43
TOTAL	1.71	1.71	1.57	2.43	**7.42**

Table 3.4: Frequencies for the verbs found in projection structures

Looking at the 'TOTAL' column first of all, the data shows that there are more verbs which potentially realise projections of ideas <44.40> than locutions <7.42>, meaning there is very little reporting or quoting of what has been said; rather, the focus is on what is thought or what is desired. There are slightly more verbs of desire <22.42> (i.e., proposals/commands requiring the recipient to act in some way) than there are verbs of cognition <21.98> (i.e. propositions/statements requiring a verbal response from the recipient). Note that HEAR, SEE, GUESS and FANCY seem to be used in a similar way to THINK (as in *what would you do if you heard Thomas and I was in Queenstown* (*JLC*, n.d. August 1895), *I see you are doing better than ever* (*JLC*, 11 May n.d.), *I gess* [*sic passim*] *she Wount* (*ELC*, 7 March 1876), *she only fancies she may be sick* (*JLC*, 9 March 1890)) and as such I would categorise these as verbs of cognition.

Focusing on each sister in turn, Table 4 shows that Lizzie (the oldest and first sister to emigrate) uses more verbs which realise propositions <7.18> than verbs which realise proposals <1.37>, THINK being the most frequent verb in Lizzie's letters, <4.11>. Going back to Table 2.2 in chapter two, which summarises the recipients of the Lough letters, it can be seen that Lizzie writes one letter addressed to her mother, two letters addressed to her mother and father and two letters addressed to her mother, father and sisters. Although there is very little data for Lizzie (just five letters in total), it is interesting to note that there are very few projections of proposals (commands) in her correspondence.

In contrast to Lizzie's letters, Alice has twice as many verbs which realise proposals <6.58> than she does verbs which realise propositions <3.43> – WISH <4.85> and HOPE <1.71> being the most common verbs realising proposals, and SUPPOSE <1.71> being the most common verb realising propositions. Unlike

Lizzie, Alice writes mainly to her sister (seven letters), with just three letters being addressed to her mother. Annie's letters show a similar pattern to Alice's letters in that there are twice as many verbs realising proposals <8.32> than there are verbs realising propositions <4.84> – with HOPE being used significantly more by Annie when compared to her sisters <7.43>. Again, Annie writes more frequently to her younger sister, Mary (26 letters), than to her mother (nine letters).

Julia's letters contain slightly more verbs which realise propositions <6.52> than verbs which realise proposals <5.88>. KNOW <1.22>, THINK <2.65> and SUPPOSE <1.94> most frequently realise propositions while HOPE <4.59> most frequently realises proposals. As Julia writes more frequently to her mother (23 letters) than to her sister (10 letters) one might expect there to be more projections of propositions (following a similar pattern to Lizzie's, Alice's and Annie's letters); however, the balance between verbs which realise propositions and verbs which realise proposals is roughly the same. A closer examination of the verbs in context is needed to see whether there is, in fact, a correlation between the use of projection structures and the relationship between author and recipient. Specifically, is the author more likely to use projections of propositions (statements) if the relationship between the interactants is unequal (i.e. children writing to parents) and is the author more likely to use projections of proposals (commands) if the relationship between the interactants is equal (i.e. letters between siblings)?

It is not within the scope of this chapter to examine all of the verbs listed in the previous section; what follows, therefore, are some general observations regarding the use of HOPE in projections of proposals and the use of THINK in projections of propositions.

Starting with Lizzie, all of Lizzie's letters are addressed to her mother. Lizzie's father and sisters are also named as addressees; however the main recipient appears to be the mother, regularly referred to using vocatives (*Dear Mother*) throughout Lizzie's letters. As mentioned previously, there are very few verbs that potentially realise projections of proposals (i.e. exchanges of goods and services – typically offers or commands) in Lizzie's letters compared with verbs that realise projections of propositions (i.e. exchanges of information – typically statements or questions). Focusing on projections of propositions containing the verb THINK, the findings show that in four (of the nine) occurrences the recipient of the letter – Lizzie's mother – is in the position of Subject of the projecting clause. In these instances, Lizzie directly involves her mother in the unfolding discourse, eliciting her views on what is being discussed (see examples 1 and 2, below). In the remaining occurrences the projection structures are used to make comments and observations about Lizzie's younger siblings – what she thought they might achieve in life and her opinions regarding their behaviour and actions (see examples 3, 4 and 5, for instance). In writing to her mother, then, Lizzie tends to use projections of propositions to seek out her mother's opinion regarding the topics being discussed or she uses them to pass comment on her younger siblings; she rarely uses projections of proposals to make (indirect) commands.

1) Elizabeth you would have had to call it Mary on account of 15 August been the blessed virgin day **dont you think it** is a prettie name Dear Mother I was very busy on account of it for the last few weeks all last week (*ELC*, 21 August 1876)
2) is so small she is not the fiste of your hand it seems as though she is not so late as she used to be **you would not think you** eaver seen her I am shure if you seen her now but still her health is pritty good **I dont think she** (*ELC*, 7 March 1876)
3) always thought she would be cute what is the matter with Maggie is she sickly or why is she no good **I always thought she** would be smart I wounder if she Wount let me no what she is best at and you did not say anything about (*ELC*, 7 March 1876)
4) time my best love to Father I am so glad he is Well my love to Mary Julia an Maggie I miss poor Nan **I think she** ought to be home I hope she is not lonsome you dont no how lonsome I feel to think she is not at home (*ELC*, 7 March 1876)
5) has evrything very nice her Husband is Well an to work evry day he earns a good pay Alice is stingy **I never thought it** untill now she is as saving as if she had a big family (*ELC*, 7 March 1876)

Figure 3.2: Concordance lines for Lizzie: THINK

In contrast to Lizzie's letters, Alice's letters contain more projections of proposals than propositions, thus requiring some action on the part of the recipient. Focusing on projections of proposals containing the verb HOPE, the data shows that, out of 18 occurrences, roughly half of these can be found in letters to Alice's younger sister Mary (eleven occurrences in total – an average of 1.57 per letter), with seven occurrences found in letters to Alice's mother (an average of 2.33 per letter). In letters to Alice's mother, there were four instances of formulaic expressions regarding the wellbeing of Alice's younger sisters, godparents and mother (see examples 1 and 2 in Figure 3.3). In the remaining three occurrences, the Subject of the projected clause is the recipient of the letter (Alice's mother). It is interesting to note that although these projection structures could be viewed as indirect commands (for instance, *I hope you wont think bad of me* (statement) can be restructured as *dont think bad of me* (command)), they do not require the recipient to physically carry out an action; rather, they require the recipient to cognitively respond in some way – by liking (example 3), thinking (example 4) or excusing (example 5), for instance. There are similar examples of projections of proposals requiring a cognitive response in Alice’s letters to her younger sister (see examples 6 and 7); however, in these letters there are also examples of projections of proposals requiring a physical response: in example 8 Alice instructs her sister to send a letter and in example 9 Alice (indirectly) instructs another family member – Maggie – to visit Mary. To summarise, then, although projections of proposals containing HOPE are found in Alice’s letters to both her mother and her sister, she is more likely to use projections of proposals (commands) that require a physical response (i.e. require the recipient to carry out an action) when writing to her younger sibling than when writing to her mother.

1) come I shall forget it Dear mother I suppose Christmas will be gone before this letter reaches you but **I hope you** will live to enjoy a good many and I wish you and Mary a happy new year I am sending you a card and (*ALC*, 18 December 1889)

2) Alicia Elliott write soon Dear Mother I suppose Christmas will be gone before the letter reaches you but **I hope you** will live to enjoy a good many and I wish you and Marry a happy New Year I am sending you a card and (*ALC*, n.d.)

3) likeness I had some taken for I wanted to send home one to you it is 14 years since I had any taken before **I hope you** will like it all though it is many years since I left home I think just as much about it as when I (*ALC*, 27 February 1888)

4) many years since I left home I think just as much about it as when I come here Dear Mother and sisters **I hope you** wont think bad of me for not writing I often think of writing but keeps putting it of from time to (*ALC*, 27 February 1888)

5) close **I hope you** will excuse all mistakes sending (*ALC*, 27 February 1888)

6) **I hope you** will forgive me for not answering sooner I have had great trouble since I wrote you last My oldest (*ALC*, 8 March n.d.)

7) honour all the family were very pleased about it and I will always remember your kindnes Dear Sister **I hope you** will excuse me for bein so long without writing but we heard so much about the war over there I dident (*ALC*, 28 December 1914)

8) Dear sister received your very welcome letter and the card was very nice I am very glad to get it and **I hope you** will send the paper you promised me I hope you will forgive me for not answering sooner I have had (*ALC*, 8 March n.d.)

9) lots of snow for xmas how is the winter over there I suppose you are bussey getting ready for xmis now **I hope Maggie** has come to see you before now and that you will have their pictures taken soon seeing I cant see your (*ALC*, 10 December, n.d.)

Figure 3.3: Concordance lines for Alice: HOPE

Similar to Alice, overall Annie's letters contain more projections of proposals than projections of propositions. Focusing on projections of proposals containing the verb HOPE, the data shows that these structures are slightly more frequent in letters addressed to Annie's younger sister Mary (104 occurrences – an average of 4 per letter) than in letters addressed to her mother (28 occurrences – an average of 3.1 per letter). There was a noticeably high frequency of these structures in the two letters sent to Annie's niece and nephew (15 occurrences – an average of 7.5 per letter). In Annie's letters, the projections of proposals more often require a physical response rather than a cognitive one. Amongst other things, Annie instructs her younger sister to write, get the home fixed and (as in example 1) keep her children at school. She also instructs her niece to remain at the family home (example 2) and her nephew to focus on his studies. Additionally, Annie uses this structure to express her desire for others to act in some way: for her nieces to visit their mother (example 3), for Maggie to visit Mary (example 4) and for Mary's children to work hard at school (example 5).

1) at home I hope they are [all?--damaged] at Deevys [tell Lizzie?--damaged] she must write to me soon **I hope you** keep them to school all you can when they grow bigger you can not send them very well I suppose there (*NLC*, 10 December 1902)
2) well. all friends here are very well and Alice I suppose you are going to school and is at home yet **I hope you** will be at home yet for a long time because it would seem lonesome if you were all gone away. I hope (*NLC*, 11 December 1914)
3) be sure and write very soon if you can **I hope some of the girls** will come that day to cheer you and (*NLC*, n.d.)
4) many things I remember about home I hope Maggie and family is very well sorry Jim lost his good friend **I hope Maggie** will come to see you often and I wish you could go see her sometimes give my love to them all when (*NLC*, 14 August 1919)
5) same age as Lizzie will graduate from the Sisters School next June I am sur Lizzie is very smart and **I hope they** will all make good use of their school days they come but once in a life time and in after years they (*NLC*, 3 April 1906)

Figure 3.4: Concordance lines for Annie: HOPE

Unlike her older sisters (Lizzie, Alice and Annie), Julia's letters contain roughly the same number of projections of propositions as projections of proposals and yet, interestingly, Julia writes more frequently to her mother (23 letters) than to her sister (10 letters). This might indicate that Julia's letters to her mother include more projections of proposals than is typically found in the letters written by her sisters. Looking more closely at projections of proposals containing the verb HOPE, there are 43 occurrences of this structure in the 23 letters sent to her mother (an average of 1.87 per letter) and 20 occurrences in the 10 letters sent to her younger sister (an average of 2 per letter). Similarly, when looking at projections of propositions that contain the verb THINK, the data shows 21 occurrences in letters addressed to her mother (an average of 0.91 per letter) and 7 occurrences in letters addressed to her sister (an average of 0.78 per letter). In other words, Julia uses roughly the same number of propositions (containing THINK) and proposals (containing HOPE) in letters to her mother and her sister.

This observation would perhaps tie in with what is known about Julia Lough. Julia was the last sister to emigrate and appears to have quickly moved up the social ladder starting off as a seamstress and ending up the proprietor of a successful dressmaking business. She is described by relatives as being 'strong-willed' and 'determined' and these traits are possibly reflected in her style of writing, in which there appears to be a lot of ego-involvement (first person pronouns). In the projections of propositions containing THINK, for instance, Julia typically uses this structure to pass comment on friends and family (see examples 1, 2 and 3 where Julia makes judgements about the behaviours and actions family members). Julia also uses the structure when reassuring and encouraging those back home (see example 4 where Julia encourages her mother to write, example 5 where she praises her sister for looking after their mother, and example 6 where she reassures her mother that she will never be forgotten). Julia also uses this structure when defending herself and/or justifying her actions: in example 7, for instance, Julia states *I think I keep you well posted* – this comment appears to be in response to an earlier letter perhaps criticising Julia for her lack of correspondence. Again, in example 8, Julia refers back to a previous letter in which, it appears, her mother and sister expressed concerns over her taking on an apprenticeship. Here, Julia directly addresses their concerns, demonstrating a confident and assertive personality.

1) a letter from Maggie last week. she seems to thing when she goes home she wont go back there again. **I think she** is very foolish now that she is getting good pay and such a good nice place I am sure she will never (*JLC*, n.d. 1889-1890)
2) the good chance you are giving them. I am glad you will have white dresses enough for Conformation [*sic passim*]. **I think Lizzie** is rather stingy about writing I hope she had a lovely visit. I am sure enjoyed Maggies visit how does (*JLC*, 24 May 1893-94)
3) see I hasen to write I am very sorry to thing [*sic passim*] Mrs [Odlunn?] got the chance to give Mag notice. **I really thought Mag** had more sense than that I think she was too well off and now to thing [*sic passim*] she has to leave when (*JLC*, 10 August 1890)
4) Mother I received all your letters. I was so surprised to get a letter in your own dear hand writing. **I think you** done just splendid it was a very nice letter and I am very thankful to you I shall always treasure (*JLC*, 9 March 1890)
5) happy in having such a good husband and Now your own children and having Mother there always but then **I think you** were always the best to Mother and it is only fair you Should receive the reward. Dear Sister we are (*JLC*, 21 March 1893)
6) sending you ten shillings so you see you are not forgotten here although Liz is a great many years here **I dont think I** would forget you either if I was away so long Indeed I never could forget my darling Mother Winsted (*JLC*, n.d. December 1888)
7) dollars up to fifty and I am sure Mrs Cleaveland pays one hundred as for letting you know all the News **I think I** keep you well posted. if I did not write but once a year I would be doing well as for Alisha I always (*JLC*, 3 September 1893)
8) take in sewing evenings as it is hard to work all the time I am very glad I made the change although **I think you and mother** did not like it by your letters of course it Seems hard to go and Work for nothing but it (*JLC*, 18 January 1891)

Figure 3.5: Concordance lines for Julia: THINK

Discussion and conclusion

This chapter began by asking how, through the language of correspondence, emigrants maintained relationships across distance and time. Previous research that explores the use of first/second person pronouns and evidential verbs in personal correspondence was used as the starting point. In line with this research, the findings of the current study showed that the Lough letters contained a high frequency of these linguistic indicators of involvement, with female authors using more first/second person pronouns and evidential verbs than their male counterparts.

The chapter then looked at these linguistic features within their wider phraseological context. Specifically, it focused on their use within projection structures, examining how these repeated patterns serve different communicative functions. By looking at the frequency of verbs that realise projections of propositions and verbs that realise projections of proposals, together with details of the author / recipient relationships, a possible correlation began to emerge. This correlation might be summarised as follows: if the author writes more frequently to a sibling, niece or nephew (an ‘inferior’ within the notional familial hierarchy) there appears to be more projections of proposals (often realising indirect commands); if the author writes more frequently to a parent (a generational ‘superior’) there appears to be more projections of propositions (typically statements, exchanging information). Additionally, letters addressed to siblings, nieces and nephews appear to contain more proposals (indirect commands) that require a physical response (i.e. they require the recipient of the letter to physically do something – send a letter, keep the children at school etc.) than letters addressed to a parent, which tend to require a cognitive response (forgive, excuse or enjoy etc.). In other words, the use of projection structures seems to reflect kinship relations: knowing one's place within the family and knowing how to write a certain way to different family members. Julia is an exception to this hypothesis, using roughly the same amount of projections of propositions and projections of proposals in letters to her mother and letters to her sister – I will be looking at Julia's letters in closer detail in chapter four.

These conclusions are, however, very tentative and further investigation is needed to see whether these initial observations hold true when looking at a much larger data set. Furthermore, the final part of this study only investigated the verbs HOPE and THINK. The verb HOPE certainly appears to be one of the strongest default verbs in personal letters to family, where the author 'performs' deference and aims to express volitive good will towards the recipient or third party; however a much more detailed study is required to see how HOPE, and other verbs, behave within projection structures. What this chapter has attempted to do, however, is to examine the function of projection structures and how they contribute to intersubjective meaning.

Both types of projection can be found in all of the Lough letters, but different types of projection will elicit different responses from the recipient. What is significant about projection structures is their ability to directly address and involve the recipient of the letter, assigning to them a role to play in the communicative event that is taking place and helping to build a psychological link between author and recipient. Projection structures may, therefore, reveal something about the relationships embodied within the letters, how the recipient is constructed and – ultimately – how family relationships are maintained. So far in this thesis I have examined the Lough correspondence at the lexicogrammatical / clause level. In chapter four, I will look more broadly at topics and themes within the letters.

CHAPTER FOUR

Case study: Identifying themes and topics within the *Julia Lough Corpus* (*JLC*)

Introduction

As discussed in the Literature Review, while the value of emigrant letters as socio-historical artefacts is now generally recognised, deciding upon the best means to exploit such resources remains problematic. Methodological issues to do with representativeness and sample size are an ongoing concern for anyone working with ego-documents, and the task of '[d]ecoding texts with inadequate paragraphing and punctuation, ungrammatical constructions [and] highly irregular spelling' – all traits typical of emigrant letters – poses various challenges with regard to content analysis (Elliott et. al. 2006, p. 4). Additionally, and perhaps most relevantly for the concerns of this chapter, there is the difficulty of what Plummer has called 'dross rate' – the fact that: 'Letters are not generally focused enough to be of analytic interest – they contain far too much material that strays from the researcher's concern' (Plummer 2001, p. 55).

The editorial practices of various scholars working on and with emigrant letters would appear to support Plummer's view. Referring to previous studies which have looked at emigrant letters from America, including those of pioneering figures like Blegen (1955) and Conway (1961) as well as more recent accounts, Fitzpatrick (1994, p. 21) has observed that: 'The authors of this otherwise exemplary work [have] shared the widespread impatience of editors with material deemed "tedious for the non-specialist," including "ritualized pious reflections" and "endless lists of persons to whom the letter-writer wishes to send his or her best regards."'[103] Consequently, what has been viewed as 'uninteresting' or 'irrelevant' material within these letters has quite often simply been omitted.[104]

Two of the most notable studies using content analysis methods to identify ideological tropes and themes within the discourse of emigrant correspondence are Miller's 1985 *Emigrants and Exiles* and Fitzpatrick's 1994 *Oceans of Consolation*. Miller examines Irish migration to North America from 1607 to 1921, arguing that although most Irish who crossed the Atlantic were 'voluntary emigrants who went abroad in search of better economic and social opportunities – that is, for the same reasons motivating emigrants from other parts of Europe' (1985, p. 6) they often viewed themselves as involuntary exiles, 'compelled to leave home by forces beyond individual control, particularly by British and landlord oppression' (1985, p. 556). To explore this incongruity, Miller analyses 5,000 emigrant letters and memoirs (as well as poems, songs, and folklore), looking at how references to homesickness and separation, as well as references to the homeland and the New World, contributed to the theme of emigration as exile. Miller's argument is that 'Irish-American homesickness, alienation, and nationalism were rooted ultimately in a traditional Irish Catholic worldview which predisposed Irish emigrants to perceive or at least justify themselves not as voluntary, ambitious emigrants but as involuntary, nonresponsible "exiles"'

[103] The quotations here are taken from Kamphoefner et. al (1988, p. 46-47).

[104] However, what I hope the previous two chapters have demonstrated is that the typical – the everyday – is in itself worthy of investigation, potentially revealing something about the various relationships embodied within the letter.

(Miller 1985, p. 556).[105] This worldview, Miller suggests, dates back to premodern times when 'Gaelic culture's secular, religious, and linguistic aspects expressed or reinforced a worldview which deemphasized and even condemned individualistic and innovative actions such as emigration' (ibid.).

Fitzpatrick, using a much smaller dataset, explores nineteenth century Irish migration to Australia. Unlike Miller, Fitzpatrick publishes his letters in full (111 letters of which 55 were sent to Australia and 56 to Ireland, between 1843 and 1906) and then analyses those letters for topics. Around 140 main themes and almost 250 sub-themes are captured in a thematic index, presented in full on pages 643-649. The index includes themes such as 'home', 'loneliness', and 'nostalgia' – features of emigrant correspondence that are also observed by Miller; however Fitzpatrick 'reports no comparable use of the ['exile' trope] among the Irish migrants in Australia' (Elliott et. al. 2006, p. 11).[106]

While Fitzpatrick's analysis of letters sent to and from Australia certainly demonstrates the benefits of a more systematised method of content analysis, both studies suggest that 'a more quantitative approach may be warranted… in order to make a better case' for the various readings and interpretations of emigrant letter collections (Elliott et al. 2006, p. 11). The present chapter then follows Miller's and Fitzpatrick's lead and uses content analysis methods to examine the emigrant letter. Like Liz Stanley, I want to argue that 'the features of letters' which others perceive 'as problems' – and dismiss as 'dross' – are 'the very things' that are 'interesting and deserving sustained attention as analytical

[105] Additionally, although to a lesser extent, the 'expulsion by force' perceived cause also led to migrant expressions of bitterness and wish for redress towards the forces that expelled them.

[106] See also O'Farrell (1984; 1987; 1990) for accounts on Irish migrants in Australia and New Zealand, based largely on letters and family memoir; and for a detailed account of patterns of Irish migration to other countries including Australia, New Zealand and South Africa see Akenson (1997).

problematics' (2010, p. 202). The everyday, the typical, and the mundane – what the emigrant wrote about over the course of weeks, years and decades – provides the fullest insight into how letters in the nineteenth century embodied, defined and modified human relationships.[107]

In order to carry out this analysis I will begin by identifying the key topics within Julia Lough's letters – the last of the Lough sisters to emigrate in 1884. I will then explore how computational methods can be used to look in more detail at the language of two of those topics in the *JLC*: 'Homesickness and Separation' and 'Recollections'. Following this, I will examine how a sense of distance as well as closeness, between author and recipient, is textually performed in Julia's letters, arguing that this helps to reinforce and 'reconfigure personal relationships made vulnerable by distance' (Elliott et. al. 2006, p. 17). Although the focus of this chapter is narrow – just 35 letters by one female emigrant – I will use this sample of material to propose a more quantifiable method of content analysis than was adopted by previous researchers, and propose also that this method may be applied more widely across other emigrant letter collections.

Methods

Of the ninety-nine letters in the Lough family collection, thirty-five are in Julia's hand,[108] and the bulk of these letters (thirty-three in total) date from 1884 to 1895. Two later surviving letters were sent from Julia to her sister between 1919

[107] Although emigrant correspondence cannot be regarded as 'the unmediated voice of pure experience', since 'immigrant writers were immersed in cultures which informed their often tentative writing' (Elliott et. al. 2006, p. 7), the content of the emigrant letter is, arguably, the best evidence available for understanding how family relationships were reinforced and reconfigured over distance and time.

[108] Note that Tables 2.1 and 2.2 in chapter two show 33 letters written by Julia. Subsequent to chapter two being written a further two letters were identified as belonging to Julia.

and 1927.[109] Julia’s primary correspondents, at least on this evidence, were her mother (who is addressed in twenty-three of the thirty-five missives) and her sister Mary (who is addressed in twelve), both of whom were still residing in Ireland. The former correspondence seems to have ceased around 1893, for in June 1894, Julia tells Mary: *I always spend the evening crying when I get a letter from you I can scarsely make up my mind yet to have Dear Mother gone you do not know how different a feeling it is to be away and try to realize what happened I have thought of Mother very much all through May* [*sic passim*]’ (*JLC*, 4 June 1894).

	Reference	From: town	From: country	Recipient	To: town	To: country	No. of Words	Day	Month	Year
1	LOUGH_005	Queenstown	Ireland	Mother	Meelick	Ireland	40	27	September	1884
2	LOUGH_062	-	Ireland / England	Mother	Meelick	Ireland	98			1884
3	LOUGH_006	Winsted	America	Sister	Meelick	Ireland	519	20	December	1884
4	LOUGH_079	Winsted	America	Mother	Meelick	Ireland	190			1884-1894
5	LOUGH_008	Winsted	America	Mother	Meelick	Ireland	342		December	1888
6	LOUGH_009	Winsted	America	Mother	Meelick	Ireland	444	3	November	1889
7	LOUGH_010	Winsted	America	Mother	Meelick	Ireland	436	2	December	1889
8	LOUGH_089	Winsted	America	Mother	Meelick	Ireland	487			1889-1890
9	LOUGH_072	Winsted	America	Mother	Meelick	Ireland	259			1889-1894
10	LOUGH_013	Winsted	America	Mother	Meelick	Ireland	463	9	March	1890
11	LOUGH_015	Winsted	America	Mother	Meelick	Ireland	366	10	August	1890
12	LOUGH_017	Winsted	America	Mother	Meelick	Ireland	350		December	1890
13	LOUGH_018	Winsted	America	Sister	Meelick	Ireland	348	18	January	1891
14	LOUGH_019	Winsted	America	Mother	Meelick	Ireland	351	25	January	1891
15	LOUGH_034	Winsted	America	Mother	Meelick	Ireland	225	30	March	1891
16	LOUGH_020	Winsted	America	Mother	Meelick	Ireland	317	18	October	1891
17	LOUGH_021	Winsted	America	Mother	Meelick	Ireland	300	14	December	1891
18	LOUGH_068	Winsted	America	Mother	Meelick	Ireland	400	11	May	pre-1892
19	LOUGH_023	Winsted	America	Mother	Meelick	Ireland	396	1	September	1892
20	LOUGH_064	Winsted	America	Mother	Meelick	Ireland	321			1892-1893
21	LOUGH_105	Winsted	America	Sister	Meelick	Ireland	423	21	March	1893
22	LOUGH_070	Winsted	America	Mother	Meelick	Ireland	305		May	1893
23	LOUGH_025	Winsted	America	Mother	Meelick	Ireland	340		July	1893
24	LOUGH_075	Winsted	America	Mother	Meelick	Ireland	356	3	September	1893
25	LOUGH_029	Winsted	America	Mother	Meelick	Ireland	334	10	October	1893
26	LOUGH_026	Winsted	America	Mother	Meelick	Ireland	451		December	1893
27	LOUGH_076	Torrington	America	Sister	Meelick	Ireland	183	25	March	1894
28	LOUGH_085	Torrington	America	Sister	Meelick	Ireland	477	24	May	1893-1894
29	LOUGH_074	Torrington	America	Sister	Meelick	Ireland	354			1889-1894
30	LOUGH_027	Winsted	America	Sister	Meelick	Ireland	736	4	June	1894
31	LOUGH_028	Winsted	America	Sister	Meelick	Ireland	469		November	1895
32	LOUGH_086	Queenstown	Ireland	Sister	Meelick	Ireland	44	8	July	1895
33	LOUGH_031	Torrington	America	Sister	Meelick	Ireland	416		August	1895
34	LOUGH_102	Torrington	America	Sister	Meelick	Ireland	331	17	March	1919-1920
35	LOUGH_057	Torrington	America	Sister	Meelick	Ireland	349	9	November	1927

Table 4.1: The *JLC*

[109] Some of the letters are not dated, but their content allows them to be placed within an approximate timeframe.

The twenty-four year gap between a letter Julia sent to Mary in August 1895 and one she sent to her in March of 1919 or 1920, meanwhile, might be explained by the fact that, in addition to managing her business these were Julia's prime childbearing years. This is not to say that Julia did not write any letters home during this period. A reference in the 1919/20 message to a *Christmas letter [being] received*, for example, suggests that the sisters did maintain some level of contact during this period, perhaps corresponding at important times of the year such as Christmas, New Year, St Patrick's Day or Easter (*JLC*, 17 March 1919/20). Yet at the same time this letter also gestures to a lack of regular contact and to a sense of estrangement as having developed between the siblings, as when Julia observes that: *We have been very well all Winter Thank God. hoping this will find You and all your family well I dont know them. You dont inform me any thing about them* [*sic passim*] (*JLC*, 17 March 1919/20). Certainly, the tone here is rather different from those earlier in the collection. The first letter we have from Julia, dated *Saturday night / September / 27-84*, was sent from Queenstown (now Cobh) on the coast of County Cork, Ireland, just before Julia embarked on her journey to America, and its very moving content gives the reader a powerful sense of what Julia was experiencing when she sat down that evening to write. *My Dearest Mother*, it begins, *but it cannot be helped now it wont* [*sic passim*] *be for long Dear Mother I would die if I thought I never would see you again you can be sure* (*JLC*, 27 September 1884). The next letter in the sequence, though it does not contain an address line or date, then seems to come just as Julia is about to set sail for America (either from Ireland or England, depending on the passage she took) and once again emphasizes the drama of departure: *I am all right so far and I hope I will sleep tonight only we and another young man at [our lodgings* *with] O. Sullivan tonight I have been at the office and we are to be out at half past seven in the morning and sail* [*sic passim*] (*JLC*, n.d. 1884).

As I have already suggested, however, concentrating on such obviously emotional passages potentially does a disservice to the full range and complexity of material in emigrant letters. To which end, I would now like to turn more specifically to the modes of discursive analysis I have been developing in an effort to broaden our comprehension of such material. In the case of Julia Lough's letters, following the assembly of digital transcriptions based on Miller's archive, my first task was to identify the key topics within Julia's correspondence through a close reading of each letter. Deciding on these topic categories was challenging in itself; more challenging still was identifying where topics began and where they ended. Indeed, the issue of topic identification is something that text-linguistics scholars have been grappling with for many years (see, for example, studies by Beaugrande (1984), Van Dijk (1977) and Hoey (1991; 2001)). According to Van Dijk, for example, 'for a sequence to have a topic, each sentence (or its underlying propositions) must "satisfy" this topic directly or indirectly' and, therefore, a change of topic can be identified 'if one of the sentences of a discourse no longer "belongs to" a given topic and if the sentence is the first member of a sequence with a different topic' (1977, p. 138). Yet as I have already indicated, such a method of analysis (which relies, to a large extent, on there being sentence boundaries within a text) is problematic when working with many emigrant letter collections. Julia's letters, for example, contain neither graphological sentences nor paragraphs. This issue is discussed by Elliott who observes that quite often emigrant letters 'struggle on', moving 'from topic to topic rarely spending time with any one matter, so that coherence is at the mercy of thematic diversity' (2006, p. 4).[110]

In order to rectify this problem, Fitzpatrick altered the formatting of the letters he included in *Oceans of Consolation* so as to 'render [them] intelligible at first reading.' (1994, p. 26). Sentence breaks were introduced 'causing changes to capitalisation and punctuation' and 'each letter [was] split into paragraphs according to topic, so laying bare at least one reading of the sequential logic of that letter' (ibid.). My approach, however, has been to not alter the formatting of Julia's correspondence; for arguably, there *is* structure and logic in her letters.[111] She certainly appears to organise her writing semantically. The rare full stops that are evident in her letters tend to indicate a change in topic, for instance (rather than the end of a sentence). Additionally too, Julia uses vocatives - such as *Dear Mother* in the June 1894 letter quoted previously - to indicate a shift in the direction of the discourse; and statements about the weather as well as references to the possibility of a reunion (as in, *I hope you and I will spend some happy time together yet*) can both indicate a topic change (*JLC*, 25 March 1894). And finally, religious references, particularly to prayers and blessing, are often used to signal the close of a particular section.

My approach, in short, has been to look for sequences within the discourse that appear to be lexically related whilst, at the same time, taking into consideration the structure and logic that already exists in Julia's letters. This is just one personal reading of the letters. From this analysis twenty-four broad categories emerged – surprisingly few perhaps given the range of possible

[110] In linguistic terms this might be described as desultory or phatic talk.

[111] And even if there is not structure and logic that too needs to be retained and studied.

subjects that might be covered by someone who is experiencing the dramatic changes that migration brings.[112] But we can take this narrow range as both a reflection of the formulaic nature of nineteenth-century correspondence, and indicative of the limited educational background of the author.[113] Tabulated in condensed form the topics (with attendant examples) are these:

Topic and Tag	**Definition**	**Example** **[*sic passim*]**
Daily Life <dailyLife>	Any descriptions of daily routines such as cooking, sewing and general household chores.	(1) 'I have been so busy getting the Sewing done before house cleaning dresses night-gowns bloomers corset-covers slip I thought I would get this done before dinner. I have leg lamb, spenach onions prune pie graham muffens, Coffee Elsie always drinks milk' (*JLC*, 17 March 1919-1920)
Death <death>	Any references to death – typically of a family member, friend or personal acquaintance -and/or mourning. Usually, such sections of a letter will be heavily loaded with religious references.	(1) 'I was very much surprised to know you did not write and let me know when dear Annie died. We could have prayed for her offered our Holy Communion's for her could have her name on the November dead list' (*JLC*, 9 November 1927) / (2) 'I was perfectly surprised to hear of Mike Fitzgerald's death how suddint. let me know did he have the priest or did he think he was going to die' (*JLC*, 25 January 1891)
Education <education>	Any mention of learning. Typically this topic only comes up when Julia is talking about her nieces and nephews.	(1) 'above all things keep they to school regular and as long as you can. There is nothing like a good education. No matter where they roam it is every thing now' (*JLC*, n.d. 1889-1894)
Enclosures <enclosure>	Any references to items that have been sent with a letter (a photograph or a newspaper clipping, for instance, or larger items such as a parcel or a trunk).	(1) 'You gave me a happy Surprise to See you drop out of the letter and I very promptly kissed you and was So glad to See you. You look well I think' (*JLC*, 17 March 1919-1920) / (2) 'I had a letter from Mary Fitzpatrick and she

[112] Although it should be remembered that these are letters to kin, not a personal diary; so all kinds of joys and sorrows that Julia experienced might not have been deemed fit to share with her mother and sister.

[113] A topic-comparison between lower-rank and higher-rank letter writers might be an interesting future study in this respect. For more on epistolary conventions in emigrant letter-writing and their relation to social background see Dossena (2007).

		sent me Maggie's picture it is very nice' (*JLC*, 2 December 1889)
Family and Friends <familyFriends>	These are often long sections, providing information, passing comment, or asking questions about family members and/or family friends. It is common for there to be several secondary topics embedded within these sections.	(1) 'Alisha was here in April. She come on Saturday and stopped till Monday' (*JLC*, 11 May pre-1892) / (2) 'I think she has a nice good place and I hope Kate Pope going there will not make any difference to her' (*JLC*, 30 March 1891) / (3) 'I am very glad he has indoor work for all winter does any of his friends ever come see him give my love to him and Mary' (*JLC*, December 1893)
Future Letters <futureLett>	Any mention of letters which are about to be written as well as letters which the author hopes to receive. Often, the author will give instructions on what the recipient should write.	(1) 'I was going to write to Mary but I will wait now till I get her next letter' (*JLC*, 9 March 1890) / (2) 'I will expect to hear from you before Christmas' (*JLC*, November 1895) / (3) 'When you write again I want to hear an account of Maggie and all particulars about home and yourself' (*JLC*, n.d. 1889-1894)
Greeting <greet>	A formulaic greeting can be found in most of Julia's letters. Variations on this greeting might include references to health or weather.	(1) 'Dear Sister we are all well here, Thank God hoping this will find you all the Same after such a cold Winter' (*JLC*, 21 March 1893) / (2) 'I am very happy to hear from you and to hear you are in the enjoyment of good health as this leaves one and all friends at present Thank God' (*JLC*, 2 December 1889)
Health and Illness <healthIll>	Any references to good health or more likely, ill health, with regard to the author, recipient, or any other persons mentioned within the content of the letter.	(1) 'if I did not take care of my self I would be often home sick it is very easy to get cold and rheumatism here and perhaps be laid up six months with it' (*JLC*, n.d. 1889-1890) / (2) 'I am so glad you did not get the grip it comes very hard on old people' (*JLC*, 9 March 1890) / (3) 'my Aunt has been very sick Since they thought she was going to die' (*JLC*, n.d. 1889-1894)
Homesickness and Separation <homeSeparation>	Any references to feelings of homesickness or separation. The word 'homesick' itself is never used in the Lough letters, but the author might refer to dreams of home, feelings of loneliness or sadness and a longing to see family members back home in Ireland. References to distance are common in this context, as are references to the passing of time and seasons.	(1) 'I was heart broken the other night I dreamed you was dead and I could not See you and you never left any message for me so I woke up crying and I was so frightened till I realized it was only a dream' (*JLC*, 25 January 1891) / (2) 'I assure you I did not cry so much in a long time as when I read your letter' (*JLC*, 25 January 1891) / (3)

		'The leaves are falling very fast today looks like winter makes me lonesome' (*JLC*, n.d. 1889-1894)
Identity <identity>	Any instances in which the author attempts to define or align themselves in some way.	(1) 'us dressmakers here are supposed to go for the latest style to those Big places and see the imported dresses' (*JLC*, 4 June 1894) / (2) 'it is not because I did not get chances to get married I am single but the right one did not come yet' (*JLC*, 3 September 1893)
Ireland and America <IrelandAmerica>	Any reference to these specific countries. Although the words 'Ireland' and 'America' are rarely used in the Lough letters, 'here' and 'there' are often used when making comparisons between home and the New World as are the pronouns 'we' and 'you.' This is often a secondary topic, such as when discussing health, work or religion.	(1) 'it is very easy to get cold and rheumatism here [in America]' (*JLC*, n.d. 1889-1890) / (2) 'I think I would hate to come back [there/to Ireland]' (*JLC*, n.d. 1889-1890) / (3) 'I think you two ought to be very comfortable there and us here working hard' (*JLC*, 9 March 1890) / (4) 'I suppose you people over there do not fast in Lent any more' (*JLC*, n.d. 1889-1894)
Migration <migration>	Any mention of emigration, whether it is the issue generally or the specific migration of family and friends.	(1) 'I was very much surprised to hear of Jim Deevy coming to this country. I hope he will like it did he ever say anything about coming to see us. … I think there ought to be a good time in Ireland soon when you say every one is coming here' (*JLC*, 11 May pre-1892)
News Event <newsEvent>	Any reporting of local, national, or international news (as opposed to family news).	(1) 'in August our dear pastor Father OKeeffe was found dead in his study. the community was shocked every one loved him' (*JLC*, 9 November 1927) / (2) 'I suppose you must have heard of the hard times is all over the country and all the shops and factories shut down We have read about some in New York Starving it seems to be a scarsety of money and all the banks have nearly all failed or closed I hope there will be some change for the better soon' (*JLC*, 3 September 1893)
Previous Letters <previousLett>	Any mention of previous correspondence, sent or received.	(1) 'Annie received your letter' (*JLC*, December 1890) / (2) 'Annie says she wrote you an ever lasting long letter last week' (*JLC*, n.d. 1889-1890) / (3) 'she wrote last week and sent you the Paper' (*JLC*, 9 March 1890)
Recollections <recollections>	There is often some overlap and ambiguity between the topic of 'Recollections' and 'Homesickness	(1) 'I hope when you go to town you go down to the chapel as you used to do

	and Separation.' Generally speaking, 'Recollections' is used when the author remembers specific events, routines, places or people from the past.	years ago and pray and spend an hour with our Lord and remember us all there' (*JLC*, 25 January 1891) / (2) 'do they still have a prosession in Convent garden. how I used to long for one of them goosebirrys' (*JLC*, 24 May 1893-1894) / (3) 'you used always be so good to me wasent I bold but I did not have common sense then have you got that little trunk yet we used to try hard to get a look in I remember' (*JLC*, n.d. 1889-1894)
Religion <religion>	Any mention of religious routines or practices (such as mass and communion), religious institutions or people, or religious discourse (this being often used to console when a death occurs within the family).	(1) 'This is a Holy day here Assension Day. Thos. went to half past five mas and I went to seven and received Holy Communion we have many devotions here three evenings a week I have not missed any so far' (*JLC*, 24 May 1893-1894) / (2) 'Father Leo is our guide and director' (*JLC*, 3 November 1889) / (3) 'Still she is only gone before us. Therefore I hope you will not fret now only be glad to think she will suffer here no more but be praying for us' (*JLC*, November 1895)
Remittance <remittance>	Any references to money being sent (or not sent) from America to Ireland.	(1) 'I am sending you one pound' (*JLC*, 3 September 1893) / (2) 'I was sorry I could not Send you and John a Xmas present, but I never gave a cent only what I Sent mother not even to Father Leo' (*JLC*, 18 January 1891) / (3) 'Dear Mother I am sending you one pound for your own especial use I am sure you want some new flannels or some thing for yourself' (*JLC*, December 1890)
Reunion <reunion>	Any instances where the author mentions the possibility that one day the family will be back together. This could be a physical reunion in Ireland or a 'heavenly reunion' after death.	(1) 'I hope I Shall meet you once more in life and have a happy time again' (*JLC*, 25 January 1891) / (2) 'With the help of god. i will see you again and I will not go all dressed in white. I am sure you would be happy to see me' (*JLC*, n.d. 1889-1890) / (3) 'I would love to see her. I hope to have that great pleasure some time in the future' (*JLC*, 1 September 1892)
Salutation <salutation>	Any formulaic opening to a letter – typically consisting of a possessive pronoun followed by the recipient's name or identifier.	(1) 'My Dear Mother' (*JLC*, 11 May pre-1892) / (2) 'My Dearest Mother' (*JLC*, 1 September 1892)

Sign Off <signOff>	Any rhetorical or structural feature marking the end of the letter. There is a sign off in most of Julia's letters, and sometimes two - one in the body and one in the margins.	(1) 'With my best love to My Dear Mother Maggie and you Dear Mary Wishing you all Mary Happy Christmas from your very Affectionate Sister Julia all sends their best love to ye hoping soon to hear from you good bye' (*JLC*, 20 December 1884) / (2) 'love to all I remain ever my dear Mother your affectionate child Julia Lough' (*JLC*, 30 March 1891)
Transportation <transportation>	Any mode of conveyance, either within Ireland or America.	(1) 'we are to be out at half past seven in the morning and sail' (*JLC*, n.d. 1884)
Weather and Seasons <weather>	Any mention of weather or the seasons. Weather seems to be an important structural feature of Julia's letters, helping to organize the discourse and to introduce or trigger other topics (especially around health and homesickness).	(1) 'The snow is about gone I hope we do not get any more. I am always glad to See beautiful Spring' (*JLC*, 21 March 1893) / (2) 'the winter was milder than usual and our spring Commenced the 21st a few patches of snow remains on the ground yet' (*JLC*, 25 March 1894)
Work <work>	Any mention of places of work, types of labor, or work routines.	(1) 'I will Soon have a trade and be more independent I work home evenings and get all the sewing I can do but when I comence to get pay I will not take in sewing evenings as it is hard to work all the time' (*JLC*, 18 January 1891)
Writing Process <writeProcess>	This category includes any references to the process of letter writing. These might include evaluative comments and statements relating to the handwriting style, neatness and spelling of the author's or recipient's writing or instances where the author describes where they are and what they are doing at the time of writing.	(1) 'I think it is about time for me to write to you' (*JLC*, 2 December 1889) / (2) 'I think you are growing smarter all the time to write such an nice letter' (*JLC*, December 1890)

Table 4.2: Topics identified within the *JLC*

As should hopefully be clear from this outline, there is some overlap between some of the topics listed in Table 4.2. 'Homesickness / Separation', 'Recollections' and 'Reunion', for instance, are certainly linked, thematically. However, there were noticeable differences between these topics which justified them having categories of their own. Whilst the topic 'Recollections' refers to instances in which Julia remembers specific events from the past (*This time a year ago she was near been called away. She used to dread the winter so much* (*JLC*, November 1895)), 'Homesickness / Separation' refers to those instances where Julia expresses feelings of nostalgia and loneliness as well as anxieties and fears about family and home (*I was heart broken the other night I dreamed you was dead and I could not See you and you never left any message for me so I woke up crying and I was so frightened till I realized it was only a dream* (*JLC*, 25 January 1891)). The topic 'Reunion', on the other hand, refers to those instances where Julia states her hope, desire or intention to, one day, return to Ireland to be reunited with her family. These tend to be short, freestanding statements, helping to reassure the recipient (Julia's mother or sister) that they are missed (*when you see me again I hope we will spent a happy time together yet perhaps sooner than you thing* [*sic passim*] *I know you would grow young again* (*JLC*, December 1888). Having identified these topic categories, the next stage was to annotate each letter accordingly. In this respect, the tags identified in angled brackets in the left-hand column of Table 4.2 were used to mark where a topic begins and where it ends. Thus for the topic 'News Event' (used to describe any reference to local, national or international incidents), the opening tag <newsEvent> was used to show where the topic started and the same closing tag with a forward slash - </newsEvent> - was used to show where the topic ended. As in the following passage from a September 1893 letter:

> **<newsEvent>**I suppose you must have heard of the hard times is all over the country and all the shops and factories shut down We have read about some in New York Starving it seems to be a scarsety of money and all the banks have nearly all failed or closed I hope there will be some change for the better soon**</newsEvent>** (*JLC*, 3 September 1893)

In cases where the discourse could be interpreted in more than one way, two or more tags were used. This meant that a section could be said to be ‘about’ a number of topics, or it could be said to be ‘about’ just one topic. In regard to the passage above, for instance, where the text is annotated with the tags for a ‘News Event,’ an alternative, or additional, interpretation might be the topic ‘Ireland and America.’ In this case the annotation would be as follows (where ‘News Event’ is the primary topic and ‘Ireland and America’ is the secondary topic):

> **<newsEvent><IrelandAmerica>**I suppose you must have heard of the hard times … I hope there will be some change for the better soon**</IrelandAmerica></newsEvent>** (*JLC*, 3 September 1893)

Additionally, it is possible for a topic (or several topics) to be embedded within a main topic. In the example below, for instance, Julia enquires about her sister’s children and as such this section could be said to be about ‘Family and Friends’. Within this section Julia makes specific reference to the importance of schooling, so the tag for the topic ‘Education’ (used to describe any mention of learning) has been embedded with ‘Family and Friends’, as follows:

> **<familyFriends>**Well Mary Dear I had no idea you had so many children I knew Lizzie was about the same age as Katherine Walsh let me know all about them and who they look like **<education>**above all things keep they to school regular and as long as you can. There is nothing like a good education. No matter where they roam it is every thing now**</education></familyFriends>** (*JLC*, n.d. 1889-1894)

Having annotated all thirty-five letters in this way it was then possible to extract all references to a particular topic – such as those dealing with the focus of this chapter – and once they were extracted, it became possible to analyse these discursive sequences using computational methods to notice lexical and grammatical patterns.

Findings

The following table lists the key topics of Julia Lough's letters in order of the number of times they occur, thereby providing us with an overview of what Julia writes about, and how often:

	Topic	Occurrences
1	Ireland and America	66
2	Family and Friends	58
3	Previous Letters	49
4	Religion	48
5	Future Letters	41
6	Greeting	35
7	Salutation	35
8	Sign Off	33
9	Weather and Seasons	31
10	Recollections	31
11	Homesickness and Separation	28
12	Health and Illness	24
13	Work	23
14	Enclosures	17
15	Remittance	16
16	News Event	10
17	Reunion	10
18	Death	9
19	Daily Life	8
20	Writing Process	8
21	Identity	6
22	Education	2
23	Migration	1
24	Transportation	1

Table 4.3: Topics in order of frequency

What is perhaps striking about this data is how rarely topics like 'Migration' and 'Transportation' seem to crop up, even though these themes are precisely the ones most researchers on emigrant letters have focused on. Instead, the far greater focus of Julia's letters is on the more subtle refraction of physical dislocation evident in the national comparisons identified under 'Ireland and America.' Other topics, meanwhile, feature heavily because they provide a recurring structure to the letters themselves, helping to organize and situate the flow of information. To see this more clearly, in Table 4.4 the topics have been separated into three columns. The Column A topics have been separated out based on their function. These tend to be highly routine and/or genre-related formulaic and structural features that occur in all correspondence[114] (that is,

[114] The term 'formulaic language' is used here to refer to multi-word units that closely resemble phrases found in similar generic points with similar functions in personal letters generally.

features which help to organise the letter content – the salutation, the greeting, references to previous and future letters, the sign off, and so on). The topics ‘Future Letters’ and ‘Previous Letters’, for instance, are a significant part of Julia’s correspondence (and emigrant correspondence more generally), often taking up large sections of the discourse, and potentially providing useful insights into letter writing networks and the flow of correspondence over time. Additionally, the ‘Greeting’, ‘Salutation’ and ‘Sign Off’ (which typically include the use of vocatives and honorifics), although very conventional and formulaic in nature, can reveal something about the educational background of the author as well as the relationship between author and recipient. Finally, ‘Weather and Seasons’ (occurring in 23 out of 35 letters) also appears to be a structural feature of Julia’s correspondence, helping to organise the discourse by signaling a change of topic. Column B shows the topics that occur 10 times or less across the 35 letters. Some of these topics (‘Daily Life’, ‘Identity’ and ‘Migration’, for instance), although not very frequent, seem to be more personal and reflexive in nature, showing moments of greatest authenticity, directness, expressiveness, and personal identity. Finally, in Column C we find the remaining – higher frequency – topics. As one might expect, the topics ‘Family and Friends’ and ‘Ireland and America’ score high in Julia’s letters. ‘Remittance’ (any reference to money sent, in this case, from America to Ireland) is a particular feature of emigrant correspondence that is certainly worth further exploration as the strategies employed by letter writers to justify and/or explain the remittance (the amount of money being sent, what it should be used for, or why money has not been sent etc.) potentially offers another layer of insight into the personal relationships embodied within the emigrant letter. One conclusion we might draw from Tables 4.3 and 4.4, then, is that emigrant letters were both self-reflexive and self-conscious in relation to the epistolary medium in which they partook.

Column A *(Topics which help to structure the letter content)*	**Column B** *(Topics with a frequency of 10 or less)*	**Column C** *(Topics with a frequency of more than 10)*
Previous Letters (49)	News Event (10)	Ireland and America (66)
Future Letters (41)	Reunion (10)	Family and Friends (58)
Greeting (35)	Death (9)	Religion (48)
Salutation (35)	Daily Life (8)	Recollections (31)
Sign Off (33)	Writing Process (8)	Homesickness / Separation (28)
Weather and Seasons (31)	Identity (6)	Health and Illness (24)
	Education (2)	Work (23)
	Migration (1)	Enclosures (17)
	Transportation (1)	Remittance (16)

Table 4.4: Topics separated into three columns

Tables 4.3 and 4.4, however, do not indicate the spread of topics across the letters as a whole. It is possible, after all, that a topic may be mentioned several times in the same letter, or it may not be mentioned at all. Counting the number of times a topic occurs thus offers only one way into Julia's letters. Counting the number of words attributed to each occurrence of a particular topic, on the other hand, arguably provides a more accurate reflection of the content of a letter, or of a letter collection. Indeed, this table looks rather different:

	Topic	Words
1	Family and Friends	4255
2	Ireland and America	2269
3	Religion	1854
4	Previous Letters	1186
5	Recollections	978
6	Work	885
7	Weather and Seasons	774
8	Homesickness and Separation	737
9	Greeting	714
10	Enclosures	699
11	Death	691
12	Future Letters	642
13	Sign Off	637
14	Health and Illness	563
15	Remittance	539
16	News Event	476
17	Daily Life	320
18	Writing Process	265
19	Identity	234
20	Reunion	213
21	Migration	136
22	Salutation	99
23	Transportation	60
24	Education	53

Table 4.5: Topics in order of word count

Here, ‘Sign Off’ and ‘Salutation,’ as one might expect given the usual brevity of these epistolary features, move down the scale from positions 7 and 8 to positions 22 and 13 respectively. Notably, however, ‘Previous Letters’ remains in the top four places (along with ‘Family and Friends,’ ‘Ireland and America,’ and ‘Religion’), thereby reaffirming the importance of the rituals and demands of letter-writing itself for Julia. Moreover, ‘Enclosures’ moves markedly up the scale – from 14th position to 10th – along with ‘Recollections,’ ‘Work,’ ‘Homesickness and Separation,’ and ‘Death’ (which rise 5 places, 8 places, 3 places and 7 places respectively). Although they do not appear as frequently as some other topics, these topics, when they do occur, are given prominence in Julia’s letters.

As mentioned previously, it is possible for a section of the letter to be ‘about’ more than one topic. In such instances, a section of the letter might be annotated several times to reflect the different interpretations. The next stage, therefore, was to see how often each topic was in primary position (i.e. where it was the main focus of a particular section of the letter) and how often it was in secondary, tertiary etc. position (i.e. where it was an alternative interpretation of a particular section, or it was a topic embedded within another, primary, topic). It should be noted that nearly all of the topics can be in primary or secondary position. However, from reading the letters, there appeared to be patterns: some topics seemed to dominate a particular letter/s whereas others seemed to be a background theme carried across all correspondence. Table 4.6 details the number of times a topic was primary (column A), or secondary, tertiary etc. (columns B to I). (The figures in Table 4.6 represent the number of occurrences of a particular topic. Focusing on 'Family / Friends', for example, we can see that this topic was in primary position (column A) 48 times, it was in secondary position (column B) 3 times, tertiary position (column C) 5 times, and so on.) Topics that most frequently occurred in primary position are shown in black; those that most frequently occurred in a secondary (or other) position (columns B to I) are highlighted in blue; and topics that tended to occur in primary and secondary position roughly the same number of times (within + or -2) are shown in red.

	Primary	*Secondary or other position*							
Topic	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Family / Friends	**48**	**3**	**5**	**1**	**1**	**0**	**0**	**0**	**0**
Salutation	**35**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Previous Letters	**34**	**12**	**1**	**0**	**1**	**1**	**0**	**0**	**0**
Greeting	**33**	**1**	**0**	**0**	**0**	**1**	**0**	**0**	**0**
Sign Off	**33**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Future Letters	**27**	**6**	**6**	**1**	**0**	**0**	**0**	**1**	**0**
Weather	**21**	**6**	**3**	**0**	**0**	**0**	**1**	**0**	**0**
Recollections	**16**	**2**	**6**	**3**	**3**	**0**	**1**	**0**	**0**
Religion	**14**	**18**	**7**	**5**	**2**	**1**	**1**	**0**	**0**
Homesickness / Separation	**13**	**9**	**5**	**0**	**0**	**0**	**1**	**0**	**0**
Remittance	**12**	**3**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
Health / Illness	**10**	**5**	**6**	**1**	**1**	**0**	**1**	**0**	**0**
News Event	**9**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
Enclosure	**8**	**4**	**3**	**2**	**0**	**0**	**0**	**0**	**0**
Reunion	**8**	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**0**
Ireland / America	**7**	**41**	**9**	**4**	**3**	**1**	**0**	**0**	**1**
Writing Process	**6**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Work	**5**	**6**	**5**	**3**	**1**	**2**	**1**	**0**	**0**
Death	**4**	**4**	**1**	**0**	**0**	**0**	**0**	**0**	**0**
Daily Life	**4**	**4**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Transportation	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Education	**0**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
Identity	**1**	**1**	**2**	**2**	**0**	**0**	**0**	**0**	**0**
Migration	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**

Table 4.6: Primary and secondary topics

Looking at Tables 4.4 and 4.6 together, some observations can be made. Structural features (those topics that are listed in Column A of Table 4.4) tend to be primary, helping to organise the flow of discourse. Topics which occur ten times or less (Column B of Table 4.4) also tend to be primary; these topics are rare, but when they do occur they are given prominence in the letters. Finally, the topics listed in Column C of Table 4.4 tend to be secondary topics; these topics – often, implicit references, repeated within and across Julia's letters – seem to contribute to underlying themes within the discourse; although these topics *can* be primary they are more often secondary. For example, looking at Table 4.6, above, the topic 'Ireland and America' is almost always in a secondary position (59 out of 66 occurrences). Although it was quite rare for Julia to speak directly about life in America (which might alienate the recipient), the reader gets a sense of her feelings, experiences and perceptions from the comments that are made in the context of other topics, such as ‘Weather’ or ‘Work’, as follows:

> **<weather><irelandAmerica>**We have had such cold rainy weather till now There is nothing planted here yet that I can see The trees are coming to bud and sweet May is here again**</irelandAmerica></weather>** (*JLC*, May 1893)

> **<work>**I am sure you work hard but Lizze will soon be able to help you work that seen good **<irelandAmerica>**We all work hard here**</irelandAmerica></work>** (*JLC*, n.d. 1889-1894)

I will now examine two of the topics in more detail: first I will explore the theme of ‘Homesickness and Separation’ and then I will look more closely at ‘Recollections’.

Homesickness and Separation

Given the particular prominence of the theme of ‘Homesickness and Separation’ in scholarship on emigrant letters we might take its rise up the rankings in Table 4.5 as simply confirming what we already know, but crucially the extraction of these discursive units also allows us to understand them in much greater detail. In this respect, all occurrences of ‘Homesickness and Separation’ having been extracted, computational tools (in this case the text corpus management and analysis system *Sketch Engine*) were used to observe distinct patterns in Julia’s language.[115] Using the 'Word List' option in 'Sketch Engine,' all the Parts of Speech (POS) tags – such as 'Preposition + Personal Pronoun' – from the *JLC* with an n-gram value of two (i.e. consisting of two words) and which occurred 10 or more times, were extracted. The following table gives the 10 most frequent of these POS 2-grams for the topic of 'Homesickness and Separation':

N-gram	**POS**	**Frequency**	**Examples (with Number of Individual Occurrences in Parentheses)**
1	*Personal Pronoun + Verb (Present Tense)*	28	I hope (5) / I wish (4) / I get (2) / I think (2) / you do (2) / you know (2) / I assure (1) / I do (1) / I go (1) / I keep (1) / I read (1) / I value (1) / I want (1) / it seem (1) / it wont (1) / we do (1) / you see (1)
2	*Preposition + Personal Pronoun*	26	for me (3) / from you (3) / of you (3) / since I (2) / so I (2) / so you (2) / till I (2) / with you (2) / between us (1) / for you (1) / if I (1) / near her (1) / near I (1) / to me (1) / to you (1)
3	*Determiner + Singular Noun*	22	the world (2) / the year (2) / a crumb (1) / a dream (1) / a feeling (1) / a letter (1) / a picture (1) / a year (1) / another baby (1) / another year (1) / any message (1) / every night (1) / every thing (1) / the evening (1) / the fall (1) / the heart (1) / the letter (1) / the office (1) / the sea (1) / the time (1)
4	*Personal Pronoun + Modal*	21	you will (5) / I would (3) / I will (3) / I could (3) / I can (2) / you can (2) / it can (1) / you could (1) / you may (1)
5	*Verb (Present Tense) + Personal Pronoun*	15	hope you (5) / wish you (3) / know I (2) / assure you (1) / get it (1) / think I (1) / want you (1) / wish I (1)
6	*Modal + Verb (Base Form)*	14	can give (1) / can hope (1) / can see (1) / could help (1) / could make (1) / could see (1) / will give (1) / will live (1) / will send (1) / will spend (1) / will write (1) / would die (1) / would enjoy (1) / would give (1)
7	*Personal Pronoun + Verb (Past Tense)*	13	I heard (2) / I thought (2) / I did (1) / I dreamed (1) / I felt (1) / I looked (1) / I made (1) / I realized (1) / I woke (1) / it seemed (1) / you gave (1)
8	*Singular Noun + Preposition*	13	time of (2) / bye for (1) / deal of (1) / home at (1) / letter from (1) / message for (1) / night since (1) / office till (1) / piece so (1) / time as (1) / year although (1) / year with (1)
9	*Adjective + Singular Noun*	13	long time (2) / beautiful spring (1) / fast today (1) / first doesnt (1) / good cry (1) / good deal (1) / grand everything (1) / little piece (1) / merry xmas (1) / next year (1) / other night (1) / other time (1)
10	*Determiner + Adjective*	13	a good (2) / a long (2) / a great (1) / a happy (1) / a little (1) / a merry (1) / all past (1) / any other (1) / the first (1) / the last (1) / the other (1)

Table 4.7: POS 2-grams occurring ten times or more in the topic 'Homesickness and Separation'

[115] Kilgarriff, A. and Kosem, I. (2012) Corpus Tools for Lexicographers. In S. Granger and M. Paquot (eds.), *Electronic Lexicography*. New York: Oxford University Press. pp. 31-56. Available from: http://www.sketchengine.co.uk. See also Kilgarriff (2004).

Looking at Table 4.7 closely, some interesting patterns begin to emerge. The n-grams ranked first and fifth, for example, contain several verbs which have the ability to project (*hope*, *wish*, *know*, *think*). Projection structures, as discussed in chapter three, consist of two main components: the project*ing* clause (*I hope*) and the project*ed* clause (*you will write*). In these structures the primary projecting clause sets up the secondary projected clause as a representation of the content either of what is thought, or of what is said (*I hope you will write*) (see Halliday and Matthiessen 2004, p. 377). Projection structures, to put it another way, have the ability to project the author's expectations, desires, or beliefs onto the recipient, or they disclose the author's expectations, desires, or beliefs to the recipient. But, as Halliday and Matthiessen have pointed out, and as discussed in chapters two and three, there is also a distinction to be made between the projection of propositions and the projection of proposals, for 'propositions, which are exchanges of information' - typically statements or questions – 'are projected mentally by processes of cognition – thinking, knowing, understanding, wondering, etc.,' whereas 'proposals, which are exchanges of goods-&-services' – typically offers or commands – 'are projected mentally by processes of desire' (2004, p. 461). In other words, whereas propositions are projected using cognitive verbs such as *know* or *think*, proposals are projected using verbs of desire, such as *hope* or *wish*. Moreover, whereas propositions prospect some kind of verbal response from the recipient (in response to the statement *I think you are growing smarter all the time*, for instance, the recipient may choose to agree or disagree), proposals prospect a non-verbal response (thus in response to the indirect command *I hope you will write*, the recipient may choose to act – by writing back – or not). These response-expecting projection structures anticipate reactions and seek to elicit certain responses from the recipient, thus contributing with particular importance to the interactive nature of letters and helping to strengthen the relationships those letters so often seek to embody.

Crucially, a closer look at the projection structures that appear in Julia's letters within the topic 'Homesickness and Separation' shows a relatively high frequency of the verbs *hope* and *wish* (see Figure 4.1). More specifically, for Julia, projection structures containing the verb *hope* are typically coupled with an expression of loneliness (as when she writes, ***I hope you will have a very happy xmas*** *Dear Mother* ***I do always be lonesome and have a good cry*** *the last three xmas*), or a statement which seeks to reassure the recipient that they are very much in her thoughts (as when she writes, ***I am thinking of you and I will not forget you*** *next year with Gods help****. I hope you will write*** *to me after Christmas*) (*JLC*, 2 December 1899, December 1890; emphasis added). Using these structures, Julia essentially does two things: she implicitly instructs her mother or sister to undertake a particular emotional or material response, whilst also reassuring them that they are an ongoing part of her mental life. In contrast to the projection structures containing the verb *hope* which prospect actions that are more or less feasible (Julia's mother and sister are required to enjoy Christmas or write a letter), those which contain the verb *wish* on the other hand typically express a desire for something which both the author and recipient know to be impossible, or at least, very unlikely – that is for Julia and her mother and sister to be physically reunited. Thus Julia variously imagines herself back in Ireland or her mother and sister to be present in America: ***I wish I was near*** *her so I could make all those things for her*; ***I wish you was near*** *so I could help you*; *Oh how* ***I wish you was near*** *... we do have every thing good* (*JLC*, July 1893, c.December

1899, 24 May 1893/94; emphasis added). Here the repetition of the word *near* across so many different letters expresses an encoded distance that textually performs a desire for closeness and shared experience.

1. **I hope you will have a very happy xmas** Dear Mother **I do always be lonesome and have a good cry** the last three xmas (*JLC*, 2 December 1889)
2. I am sure you will be just as well pleased when you know **I am thinking of you and I will not forget you** next year with Gods help. **I hope you will write** to me after Christmas. **I hope you will spend a very happy xmas** Mother Dear and that you will live to see a great many more you may be sure **I am thinking of you** although the sea rools between us **I am always lonesome** for you but more at xmas than any other time. (*JLC*, December 1890)
3. Christmas is all past and gone **I thought very much of you** and what change a year brought to you. **I hope you was very happy** and spent a merry Xmas (*JLC*,18 January 1891)
4. **I wish I was near** her so I could make all those things for her (*JLC*, July 1893)
5. **I wish you was near** so I could help you (*JLC*, n.d. 1889-1894)
6. Oh how **I wish you was near** I made cake and pies puddings biscuits jellies we do have every thing good (*JLC*,24 May 1893-1894)
7. **I wish you could see it** I think I will send you a little piece so you can see what I am wearing (*JLC*, n.d. 1884-1894)

Figure 4.1: Sample concordance lines for n-gram 1 of Table 4.7

Interestingly, these kinds of projection structures often contain modal verbs (as in, ***I hope** you **will** spend a very happy xmas Mother Dear*), but it is the modals that do *not* belong to these structures that are most often used to construct hypothetical worlds that function to create a sense of closeness and immediacy between author and recipient (*JLC*, n.d. December 1890; emphasis added) – see Figure 4.2. Looking at occurrences of the modal *would*, for example, in writing *Dear Mother **I would die** if I thought I never would see you again you can be sure* Julia imagines a world in which she and her mother *will* meet again, and in writing ***I would give the world to be with you tonight** but do not fret I would not like to say good bye for long* she imagines a world in which her return to Ireland *is* possible (n.d. 1884; emphasis added).[116] Equally, the modal verb *will* – typically used to reassure Julia's mother that she is going to be permanently missed and remembered – seems to function in a similar way. Thus, when she writes *Dear Mother I want you to be very happy this xmas and …* ***I will be with you*** *all for home is where the heart is* or *I am sure you will be just as well pleased when you know I am thinking of you and* ***I will not forget you next year with Gods help***, rather than speaking in the present tense, Julia conspicuously shifts her attention to the future (*JLC*, December 1893, December 1890; emphasis added). This deictic shifting between worlds, as realised through references to person (I/you), time (past/future), distance (near/far) and location (here/there), both performs and reinforces homesickness and separation whilst at the same time creating what Fitzpatrick has called 'common moments of imaginable communion' (1994, p. 494). Or as Julia herself succinctly puts it: *I am thinking of you although the sea rools [sic passim] between us* (*JLC*, December 1890).

1. Dear Mother **I would die** if I thought I never would see you again you can be sure (*JLC*,27 September 1884)
2. **I would give the world to be with you tonight** but do not fret I would not like to say good bye for long (*JLC*, n.d. 1884)
3. how **I would enjoy being home** and see how grand everything looks there but I can hope (*JLC*, 21 March 1893)
4. Now my Dear Mother I want you to be very happy this xmas and do not fret about anything only say your prayers with all your heart and please remember me **I will be with you** all for home is where the heart is (*JLC*, December 1893)
5. I am sure you will be just as well pleased when you know I am thinking of you and **I will not forget you next year with Gods help.** I hope you will write to me after Christmas. I hope you will spend a very happy xmas Mother Dear and that you will live to see a great many more you may be sure I am thinking of you although the sea rools between us I am always lonesome for you but more at xmas than any other time (*JLC*, December 1890)

Figure 4.2: Sample concordance lines for n-gram 4 of Table 4.7

[116] These two examples, 'would die' and 'would give the world', are perhaps formulaic alternatives to expressions of strong desire/wishing.

Finally, a closer look at the n-grams derived from the last three POS identifiers in Table 4.7 confirms the importance of the temporal markers (*xmas*, *tonight*, *next year*) evident in the passages just quoted to Julia's discursive world. The following are examples of 'Singular Nouns and Prepositions' (SNP), 'Adjectives and Singular Nouns' (ASN), and 'Determiners and Adjectives' (DA), in context (with emphasis added):

SNP: *the fall is the nicest* ***time of*** *the year here but it always makes me lonesome* (*JLC*, 18 October 1891)

SNP: *I think this is the nicest* ***time of*** *the year although lonesome* (*JLC*, 10 October 1893)

ASN: *you know I am thinking of you and I will not forget you* ***next year*** *with Gods help* (*JLC*, December 1890)

ASN: *I am always glad to See [sic passim]* ***beautiful Spring****, how I would enjoy being home and see how grand everything looks there* (*JLC*, 21 March 1893)

DA: *I assure you I did not cry so much in* ***a long*** *time as when I read your letter* (*JLC*, 10 August 1890)

DA: *Christmas is* ***all past*** *and gone I thought very much of you* (*JLC*, 11 May c.1892)

What is striking here is how the three lexical sets are united by the recurrence and interrelation of three particular topics: 'Homesickness and Separation,' 'Recollections,' and 'Weather and Seasons.' Once again, the key phrases that *Sketch Engine* has extracted from the letters serve a bridging function, joining people, periods and places into imagined wholes. But in this instance we can see more clearly how changes in the season or climate seem to have triggered feelings of distance and longing in Julia. It is not always the case that the n-grams point to reflections on the relationship between past and present, however, as the equal frequency of 'Personal Pronouns and Verbs (Past Tense)' in Table 4.7 might indicate. Importantly, a closer look at these past tense verbs in context not only reveals that they tend to be verbs of perception or cognition, but that they are perception/cognition verbs which suggest a lack of clarity (see Figure 4.3, below).

1. I was heart broken the other night **I dreamed** you was dead and I could not See you and you never left any message for me so **I woke up** crying and I was so frightened till **I realized** it was only a dream (*JLC*, 25 January 1891)
2. **It seemed** so long since I heard from home I was getting uneasy (*JLC*, 10 August 1890)
3. so you do not know how much **I felt** when I looked upon your face again if only in a picture (*JLC*, 4 June 1894)

Figure 4.3: Sample concordance lines for n-gram 7 of Table 4.7

Thus, in one letter Julia states that ***It seemed*** *so long since I heard from home I was getting uneasy*, while in another she writes, ***I dreamed*** *you was dead and I could not See* [*sic passim*] *you and you never left any message for me so* ***I woke up*** *crying and I was so frightened till* ***I realized*** *it was only a dream* (*JLC*, 10 August 1890, 25 January 1891; emphasis added). Within these discursive structures, then, there is an overwhelming sense of vagueness and uncertainty – a lack of knowledge about Ireland, friends and family – that suggests that the imaginative bridging of physical distance was not always so easily achieved for Julia. This sense of 'not-knowing' was also evident when I used the online corpus analysis and comparison tool *Wmatrix*[117] to identify key semantic fields within 'Homesickness and Separation'. The results are summarised in Table 4.8, below, and show that the semantic fields 'Seem' and 'Mental actions and processes' (semantic fields which include words such as 'seems', 'looks' and 'dreamed') are statistically significant in 'Homesickness and Separation' when compared against a general reference corpus of letters.

Semantic Tag	**Semantic Field**	**LL Score**	**Examples**
N6+++	Frequent	78.27	always
N6---	Quantities: little	31.83	lonesome
O4.2+++	Judgment of appearance	18.25	nicest
Z8	Pronouns	12.71	I, you, it, me, your, what, she, we, that, her, this, ye, us, my, anything, which everything
A8	Seem	12.19	Seems, seemed, looked, seem, looks
E4.1+	Happy	11.64	Happy, merry
X3.4	Sensory: Sight	10.76	see
S9	Religion and the supernatural	10.10	xmas, Christmas, God, gods, prayers, prayed
X2	Mental actions and processes	8.53	dreamed
A7+	Likely	6.84	Would, can, sure, could, assure, may

Table 4.8: Key semantic fields in 'Homesickness and Separation' compared with a reference corpus

Recollections

Looking now at the topic 'Recollections', the same procedure was followed to extract POS 2-grams which occurred ten or more times. What follows is a summary of the main observations.

[117] Rayson, P. (2009) *Wmatrix*. Lancaster University. Available from: http://ucrel.lancs.ac.uk/wmatrix/ [Accessed 1 June 2011]. See also Rayson (2008).

N-gram	POS	Frequency	Examples (with Number of Individual Occurrences in Parentheses)
1	*Personal Pronoun + Verb (Present Tense)*	30	I remember (7) / I hope (6) / I suppose (5) / I know (2) / you see (2) / I dont (1) / I think (1) / me know (1) / you do (1) / you hate (1) / you know (1) / you look (1) / you make (1)
2	*Determiner + Singular Noun*	23	every thing (3) / the family (2) / a look (1) / a picture (1) / a prosession (1) / a shilling (1) / a year (1) / all winter (1) / an try (1) / any way (1) / every night (1) / that yard (1) / the fall (1) / the last (1) / the rest (1) / the side (1) / the size (1) / the winter (1) / this time (2)
3	*Personal Pronoun + Verb (Past Tense)*	22	she used (3) / we used (3) / you used (3) / I used (2) / I did (2) / he used (1) / I looked (1) / I noticed (1) / I saw (1) / it recalled (1) / it used (1) / you got (1) / you reminded (1) / I felt (1)
4	*Preposition + Personal Pronoun*	21	of them (2) / of you (2) / to me (2) / to you (2) / with me (1) / about me (1) / about us (1) / as it (1) / as you (1) / for me (1) / for us (1) / if I (1) / if it (1) / if she (1) / in I (1) / like me (1) / to we (1)
5	*Infinitive 'To' + Verb (Base Form)*	19	to get (3) / to go (2) / to mass (2) / to buy (1) / to do (1) / to dread (1) / to give (1) / to hear (1) / to pray (1) / to promise (1) / to say (1) / to see (1) / to think (1) / to try (1) / to write (1)
6	*Personal Pronoun + Adverb*	19	I never (2) / I often (2) / she always (2) / I always (1) / me not (1) / she often (1) / them well (1) / they still (1) / us all (1) / us here (1) / you either (1) / you good (1) / you often (1) / you so (1) / you still (1) / you yet (1)
7	*Adjective + Singular Noun*	17	back view (1) / common sense (1) / convent garden (1) / different life (1) / fine time (1) / good picture (1) / good time (1) great change (1) hearty cry (1) / last evening (1) / last time (1) / last year (1) / little trunk (1) / other night (1) / poor picture (1) / precious baby (1) / red ribbon (1)
8	*Verb (Present Tense) + Personal Pronoun*	15	suppose you (4) / hope you (3) / do they (1) / hope she (1) / know I (1) / know you (1) / remember them (1) / remember you (1) / see you (1) / suppose he (1)
9	*Adverb + Adjective*	14	as bad (2) / very thankful (2) / almost past (1) / as good (1) / just right (1) / not able (1) / only last (1) / so good (1) / so many (1) / so much (1) / very happy (1) / very poor (1)
10	*Determiner + Adjective*	14	a great (4) / the same (3) / a good (2) / a different (1) / that precious (1) / the back (1) / the last (1) / the other (1)

Table 4.9: POS 2-grams occurring ten times or more in the topic 'Recollections'

Looking at Table 4.9, n-gram 1 shows a relatively high frequency of the pattern *Personal Pronoun + (Present Tense) Verb* in Julia's letters, as in *I remember*, *I hope*, *I suppose* etc. Perhaps unsurprisingly, *I/she/you remember* is the most frequent combination (see examples (1) to (8) in Figure 4.4). In six of these occurrences, Julia is the subject of the clause – the participant who is remembering. Julia remembers physical objects: for instance, *those beads* (3); actions and events: *how long you used to pray* (5); and experiences and feelings: *how delighted I was* (4). The act of remembering is evident across most of the letters and serves to authenticate Julia's attachment to the homeland. Recalling specific details about people, places and events creates a bridge between the two worlds enabling author and recipient to be united through their past, shared experiences.

Turning now to n-gram 3 – *Personal Pronoun + (Past Tense) Verb*, as in *I used to*, *I looked*, *I noticed* etc. – 13 out of the 22 occurrences contain the verb *used (to)*. Julia is the subject of just two of those structures: *how delighted I used to be* (4), and *I used to long for one of them goosebirrys* [*sic passim*] (12). In the remaining 11 occurrences, Julia recalls the actions, routines and habits of others: her father (4), her family (5, 6), her sister (5, 6, 9, 13), and her mother (10, 11). In these occurrences, Julia reassures the recipient of the letter (her mother or sister) that they – and family in Ireland – are remembered, whilst at the same time demonstrating that she knows and understands their likes, dislikes, traits, fears and routines. In these occurrences a sense of knowing is textually performed through the language of recollection. By writing about shared experiences and by demonstrating that she remembers all details about home, Julia seeks to reinforce bonds with loved ones in Ireland.

1. Liz is very thankful to you **She** often talks about home and **remembers** every thing that happened there (*JLC*, 3 November 1889)
2. write so and dont forget that where ever I am **I remember** you yet (*JLC*, 1 September 1892)
3. supposing you are not able to go to mass all winter I am sure you make those beads of yours rattle in fine time every night. **I remember** them well the size of them (*JLC*, December 1893)
4. **I remember** how delighted I used to be when he used to give me a shilling at xmas I hope Dear Father is praying for us all in Heaven. (*JLC*, December 1893)
5. I have thought of Mother very much all through May **I remember** the prayers **we used to** say during May let me know do you pray as much now as when I was at home. **I remember** well how long **you used to** pray I never could be as good as you any way Dear Sister I often think of those Dear old happy days. I know **you used to** always agree with me in everything an try to think every thing I did was right (*JLC*, 4 June 1894)
6. **you used** always be so good to me wasent I bold but I did not have common sense then have you got that little trunk yet we used to try hard to get a look in **I remember** (*JLC*, n.d. 1889-1894)
7. That is certainly the back view of Asylum. in viewing it it recalled a great many things to my mind **I remember** going there to see Father (*JLC*, 24 May 1893-1894)
8. you see how little I know after so many years, **you** will **remember** us (*JLC*, 17 March 1919-1920)
9. I am sure you are not lonesome with that precious baby if she looks like me Mary will have to buy her that yard of red ribbon **she used to** promise me (*JLC*, 1 September 1892)
10. This time a year ago she was near been called away. **She used to** dread the winter so much (*JLC*, November 1895)
11. does your cough be as bad as **it used to** be or do you go to Mass every Sunday I hope you get along well (*JLC*, 30 March 1891)
12. do they still have a prosession in Convent garden. how **I used to** long for one of them goosebirrys (*JLC*, 24 May 1893-1894)
13. I often think of Mary when **she used to** go to [Toyer?] I hope she has a good time now and sleeps till nine oclock mornings (*JLC*, n.d. 1889-1890)
14. Indeed I never could forget my darling Mother (*JLC*, December 1888)

Figure 4.4: Sample concordance lines for n-grams 1 and 3 of Table 4.9

N-gram 6 shows a relatively high frequency of the pattern *Personal Pronoun + Adverb*, as in *I never*, *I often*, *she always* etc. Some of these pronoun/adverb combinations can be found in the examples above (underlined in examples 1, 5, 6, 12 and 13). Here, the adverbs are used to emphasise the frequency with which Julia thinks about home: *I often think of those Dear old happy days* (5), and *I often think of Mary* (13), for instance. In the case of example (14) – *I never could forget my darling Mother* – Julia emphasises the impossibility of her ever being able to forget. Additionally, the adverb *always* seems to be used to emphasise the sense of knowing mentioned previously. In examples (5) and (6) Julia demonstrates that she knows and understands her sister based on past experiences, repeated over time: *you used to always agree with me*, *you used always be so good to me*. This demonstration of knowledge about family seems to be a strategy for reinforcing family bonds.

Another observation, looking at examples (1) to (14) above, is to do with the use of time deixis including seasons: *Winter* (3, 10), months *May* (5), and yearly events *Xmas* (4), as well as references to the passing of time: *after so many years* (8), and *this time a year ago* (10). Seasons, months and yearly events appear to trigger specific memories. These deictic features place Julia at a particular point in time: writing in the present, she places herself firmly in the past to a period when she and her family were together.

N-gram 8 shows a relatively high frequency of the pattern *Present Tense Verb + Personal Pronoun*, as in *suppose you*, *hope you*, *know you* etc. In structures containing the verb *suppose*, Julia is always the subject of the projecting clause (*I suppose*). These structures seem to function in two ways: 1) they contribute to the interactive nature of the letters, requiring the recipient to agree, disagree, confirm or deny the statements being put forward; 2) they help to construct an imagined world based on Julia's past knowledge of family and friends in Ireland. This imagined homeland relies, however, on things in Ireland having stayed the same since Julia's departure: *supposing you are not able to go to mass* (15), *I suppose you still do the same* (18), and *I suppose you look about the same* (19). In these occurrences Julia predicts that people, places and routines have not changed in Ireland – people do the same, and look the same. Unfortunately, letters from Julia's mother and sister are not available, so the extent to which Julia's family in Ireland confirmed or rejected these projections is unknown. In summary, the verb *suppose* is very ‘Other’ oriented. In using *I suppose you* the author performs awareness of the recipient’s world (i.e. the content of the projected clause is about the recipient’s world rather than the author’s). In imagining the recipient’s world the author shows how vivid that world – home – is for them.

In contrast, the verb ‘hope’ seems to represent powerless wishing – it is a very deferential verb. It expresses a wish for another person without assuming the right or the power to make the wish come true. In some ways it resembles praying to a greater power – the author hopes or wishes for things for other people without making any presumption that they have the right, power or authority about whether it happens, or not, as in: *you are much smarter than when I was home but I hope you will not have so much to do anymore* (*JLC*, n.d. December 1888).

15. **supposing** you are not able to go to mass all winter (*JLC*, December 1893)
16. I was dreaming the other night about Dick Conroy. I suppose he is married by this time (*JLC*, November 1895)
17. **I suppose** you often talk about me and what a little snit I was but you know I am eight years older now and that makes a great change we will hope for the better it was only last evening Liz and I was talking she says she always considered me different in all my ways from the rest of the family. I think every thing she says is just right - she always cared for me and treated me the best in the family (*JLC*, 10 October 1893)
18. **I suppose** you still do the same with yours (*JLC*, 24 May 1893-1894)
19. **I suppose** you look about the same you See what a different life yours and mine has been (*JLC*, 21 March 1893)

Figure 4.5: Sample concordance lines for n-gram 1 of Table 4.9

Finally, n-gram 9 shows a relatively high frequency of the pattern *Adverb + Adjective*, as in ‘as bad’; and in n-gram 10 we see a relatively high frequency of the pattern *Determiner + Adjective*, as in *the same*. In examples (20) and (21) – see Figure 6 – Julia appears to be posing questions relating to her mother’s health and whether her health is the same, or different (that is, worse). Examples (22) and (23) are part of the projection structures mentioned previously. In example (23), however, Julia gestures to a sense of difference. Whilst Julia predicts that things are the same in Ireland, she suggests that things are very different for her in America. And in example (24) Julia reports that she has changed. Ireland represents sameness, a lack of change, while America represents progress.

20. does your cough be **as bad as** it used to be or do you go to Mass every Sunday I hope you get along well (*JLC*,30 March 1891)
21. and if your Cough does be **as bad as** usual and are you able to get out to mass every Sunday (*JLC*, 25 January 1891)
22. I suppose you still do **the same** with yours (*JLC*, 24 May 1893-1894)
23. I suppose you look about **the same** you See what **a different** life yours and mine has been (*JLC*, 21 March 1893)
24. but you know I am eight years older now and that makes **a great change** (*JLC*, 10 October 1893)

Figure 4.6: Sample concordance lines for n-grams 9 and 10 of Table 4.9

The *Wmatrix* results for key semantic fields in 'Recollections' certainly seem to support the observations discussed so far with the semantic fields 'Knowledgeable' / 'No knowledge' and 'Thought, belief' drawing our attention to verbs which express memories (*remember*, *used to* etc.) and predictions (*suppose*) – see Table 10. However, some new observations do appear to come to light: the semantic fields 'Happy' and 'Evaluation: Good' may suggest a connection between recollections and positive emotions, as in: *Dear Sister I often think of those Dear old happy days* (*JLC*, 4 June 1894) and *I remember how delighted I used to be when he gave me a shilling at xmas* (*JLC*, n.d. December 1893). Recollections, it seems, evoke positive feelings.

Semantic Tag	Semantic Field	LL Score	Examples
T1.1.1	Time: Past	40.12	used to, last year, last time, ago, the other night, last evening
X2.2+	Knowledgeable	18.48	remember, now, remembers, recalled
X2.1	Thought, belief	15.67	think, suppose, considered, trust, thought, felt, thinking, viewing
N6+++	Frequent	14.79	always
X2.2-	No knowledge	13.12	forget, forgotten
N6+	Frequent	12.35	often, every Sunday, again, every night
E4.1+	Happy	9.83	happy, joys, enjoy yourself, delighted
Z8	Pronouns	9.08	I, you me, she, it, that, your, we, us, yours, them, my, he, they, what, its, her, everything
A8	Seem	7.18	looked, look
A5.1+	Evaluation: Good	6.82	good, great, well, fine

Table 4.10: Key semantic fields in 'Recollections' compared with a reference corpus

Discussion and conclusions

Ultimately, the methodologies I have been outlining in this chapter allow us to see how Julia's relationships were changed, maintained, constructed and performed through language. With particular regard to the topic of 'Homesickness and Separation', projection structures are used to anticipate responses and reactions, assigning the recipient of the letter a role to play in the unfolding discourse; modal verbs are used to help to construct possible worlds in which the author and recipient might once again be reunited; social deixis textually constructs a sense of distance and separation between participants; and finally lexis relating to loneliness and sadness, as well as cognition and perception verbs expressing vagueness, all contribute to what might be described as a lexicogrammar of emigrant epistolarity.

With regard to the topic 'Recollections' the verbs *remember* and *used (to)* feature heavily in Julia's letters. Julia recalls the personal traits and physical appearances of family back in Ireland (what they used to do and what they used to look like, for instance) as well as remembering specific places, events, and experiences. This contributes to a theme of knowing within Julia's letters. By recounting, in very specific detail, a person, place or event, Julia is able to connect with family back home. As mentioned previously in this thesis, ‘home’, as Fitzpatrick puts it, becomes ‘a spiritual rendezvous for separated kinsfolk’ providing correspondents with ‘common moments of imaginable communion’ (1994, p. 494). Additionally, *Personal Pronoun + Adverb* combinations (*I always*, *I often*) are used to emphasise how frequently Julia remembers home, while time diexis (references to months, seasons and annual celebrations, such as Christmas or St Patrick’s Day) are a trigger for certain memories and, it would seem, positive emotions. Another significant feature of the language of recollections is the high frequency of projection structures containing the verb *suppose*, which are used to construct an imagined homeland based on past, shared experience. In these structures Julia predicts that things have not changed in Ireland – the landscape, the people and places are exactly as Julia left them. In contrast, however, America represents change, difference and progress. This dualistic position is, arguably, a common feature of emigrant letters more generally where ‘the greater the tensions incidental to exposure to new social systems and cultures, the greater…the desire to preserve a feeling of rootedness in a personal past’ (Elliott et. al. 2006, p. 2). The idea of Ireland representing sameness, is, perhaps, central to Julia’s need for ‘rootedness’ and as such it is imposed onto the recipient (Julia’s mother and sister) as without this common ground Julia’s sense of self, in relation to her family, may be threatened. As mentioned previously, unfortunately, within the Lough collection, letters from the homeland are not represented. However, one might conjecture that it is unlikely that Julia’s mother would be making similar predictions about what Julia is doing on a day-to-day basis, given that her understanding of American life would have been based purely on what Julia (and her sisters) wrote in their letters. In other words, while Julia can (and frequently does) predict the routines of family in Ireland, it is not possible for Julia's mother to do the same.

As already noted, the focus of this chapter has been narrow, examining just one collection of correspondence – 35 letters by one Irish emigrant to America – from Miller's much larger archive of 5,000 emigrant letters. But through repeating the process I have described here, using letters by authors from a range of socio-historical, economic and cultural backgrounds, a more comprehensive lexicogrammar of my key topics may begin to emerge, providing a fuller picture of the language and functions of emigrant correspondence, whilst also, potentially, paving the way for semi-automated methods of topic identification for emigrant letter collections in the future. By analysing the language of each topic – identifying words, phrases and patterns that are instances of the thematisation of, for instance, homesickness and separation – it may be possible to identify a range of local grammars, which, in turn, can be used to identify topics in other letter collections. Equally too, of course, this further research may show that the linguistic features and themes I have identified here need to be expanded or refined as other, more typical ones emerge. Nor should we forget that the discourses and topics that do not emerge may be as telling as the ones that do. Indeed, a keyword or key semantic tag comparison with a suitable reference corpus might pinpoint some notable absences, as negative key items, for instance.

Certainly, from reading Julia's letters, one gets the feeling that her emigration operated as a great source of guilt and regret. In a letter to her sister from 1893, for example, she declares, *See what a different life yours and mine* *has been I am sure you are happy in having such a good husband and Now your own children and having Mother there always but then I think you were always the best to Mother and it is only fair you Should receive the reward* [*sic passim*] (*JLC*, 21 March 1893). Yet, by all accounts, Julia's life in America was very successful and prosperous; she had independence, a career, a business and a family. It may simply be the case that Julia feels socially constrained by nineteenth-century attitudes about Irish emigration and the emigrant experience here, and that she does not want to offend her sister by emphasising the positive possibilities of the New World. It is striking, though, that as well as rarely mentioning her own work and family in her letters, Julia at no point states that she is happy in America. In this respect, what is *not* talked about in emigrant letters may be just as interesting, and as revealing, as what *is* talked about.

CHAPTER FIVE

The digitisation, annotation and visualisation of emigrant letter collections: describing and organising metadata within the TEI header

Introduction

In previous chapters I have looked at how corpus tools can be used to explore the content of digitised emigrant letters. An important next step is to bring together similar but at the same time various, different and separate digital letter collection, thereby allowing these types of analyses to be carried out across much larger bodies of data.

Many existing digital emigrant letter collections consist of unannotated versions of original manuscripts. The digitisation process has made the letters more accessible to academics and the general public, and has also increased their searchability, at least to a certain extent. Unfortunately, however, emigrant correspondence projects have often evolved independently of one another, and although project teams have been successful in tackling important research questions relating to social history and immigration studies, relatively few projects have moved beyond the digitisation stage to exploit text content and enhance usability and searchability through the use of corpus techniques[118] and visualisation tools. Different emigrant letter collections cannot easily interconnect if they are simply digitised without markup, and some search pathways through

[118] Corpus techniques of analysis involve using corpus tools such as *AntConc*, *Sketch Engine* and *Wmatrix* to extract and organise wordlists, n-grams, clusters, type/token ratios, collocations and so on (as demonstrated in chapters two to four).

the material will remain unavailable if software tools are not employed to process this encoding.

This chapter will demonstrate how interdisciplinary emigrant letter edition projects could benefit from a consistent TEI (Text Encoding Initiative) encoding, allowing correspondence metadata to be described, organised, managed and visualised in potentially useful ways. It will also look at how the findings from chapters one to four might be incorporated into a system of markup for emigrant correspondence, thereby allowing letters to be searched by topics, themes and other linguistic information.

The chapter begins with a brief introduction to encoding. It then explains the different stages of the encoding process, from document analysis to text markup, before providing examples of how visualisation tools might be used to explore correspondence metadata in useful ways.

The encoding process

Figure 5.1 outlines the encoding process, which begins with a close analysis of the research material at hand (circled on the left): the original manuscript (hereon in referred to as the document) and/or the transcription of that original manuscript (hereon in referred to as the text). I say 'and/or' because, quite often, the researcher only has access to a transcription, as the original manuscript may no longer be accessible. The Lough family letters, referred to throughout this thesis, are a case in point.

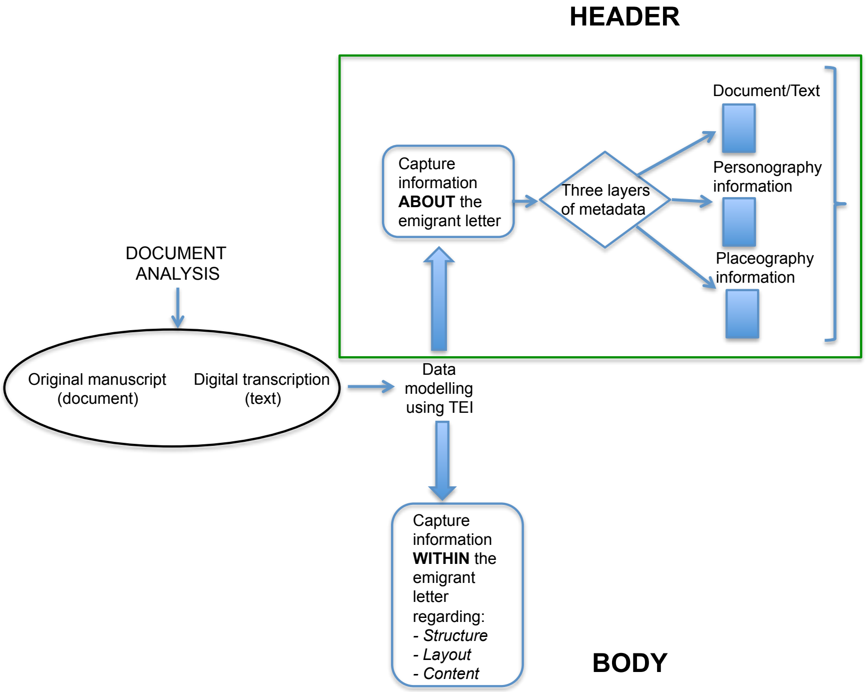
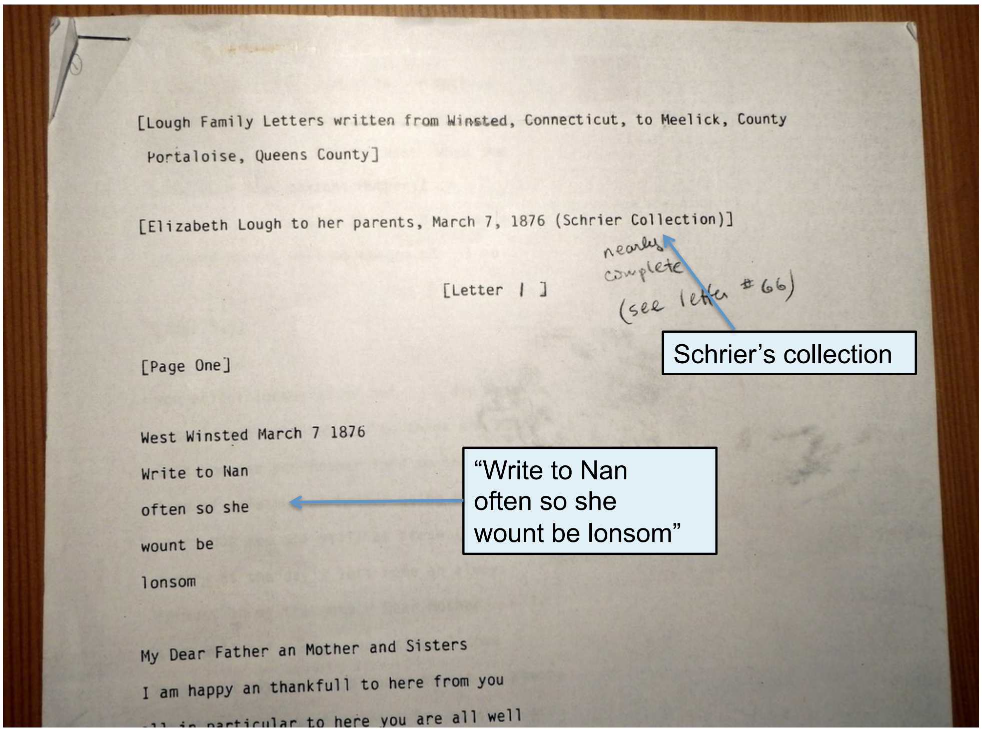

HEADER
Document/Text
Capture information ABOUT the emigrant letter
Three layers of metadata
Personography information
Placeography information
DOCUMENT ANALYSIS
Original manuscript (document)
Digital transcription (text)
Data modelling using TEI
Capture information WITHIN the emigrant letter regarding:
- Structure
- Layout
- Content
BODY

Figure 5.1: The encoding process

As outlined in chapter one, there are 99 letters in the Lough collection. These letters are drawn from a much larger body of Irish emigrant correspondence that has been collected by Professor Miller and that is housed at the University of Missouri. In the 1950s a few of the Lough letters were donated to Arnold Schrier (Professor Emeritus at the University of Cincinnati). Schrier made transcriptions of the letters and returned the original manuscripts to the donors. Later, in the 1970s and 1980s, the rest of the Lough letters were donated to Miller. Like Schrier, Miller made transcriptions and returned the original manuscripts to the donors. The Lough collection, therefore, contains Miller's and Schrier's typed transcriptions, together with a small number of photocopies of the original manuscripts.

Working with transcriptions, without having access to the original manuscript, can lead to various problems – especially if the transcription practices have not been documented in a formalised way (as is the case here). In instances where there are several transcriptions of the same letter, for example, it is not always clear which version is the final, and most reliable, copy. Additionally, different information is transcribed in different ways by Schrier and Miller, making it difficult to interpret what various textual annotations (such as [square brackets], -dashes-, or *asterisks*) might represent. It is possible to demonstrate some of these issues by looking at one of the letters in the Lough collection: a letter by Elizabeth Lough – the oldest of the Lough sisters and the first to emigrate in around 1870-1871. The letter was written on 7 March 1876 and is addressed to Elizabeth's parents and sisters in Meelick, Ireland. Figure 5.2 is Schrier's typed transcription of Elizabeth's letter. Later, an additional four pages of this letter were discovered and transcribed by Miller (Figure 5.3). Miller then pieced these fragments together to produce a complete version of the letter (Figure 5.4). Finally, Miller's research assistant produced a digital transcription of the complete letter in MS Word (Figure 5.5).



1

[Lough Family Letters written from Winsted, Connecticut, to Meelick, County Portaloise, Queens County]

[Elizabeth Lough to her parents, March 7, 1876 (Schrier Collection)]

nearly complete (see letter #66)

[Letter 1]

Schrier's collection

[Page One]

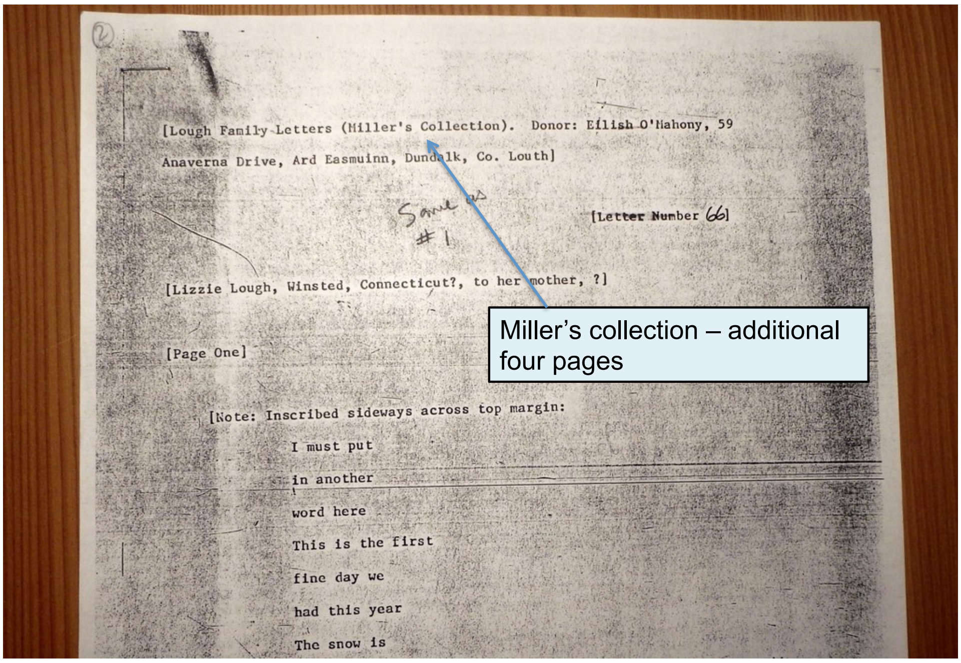
West Winsted March 7 1876

Write to Nan
often so she
wount be
lonsom

"Write to Nan often so she wount be lonsom"

My Dear Father an Mother and Sisters
I am happy an thankfull to here from you
all in particular to here you are all well

Figure 5.2: Schrier's transcription of a letter by Elizabeth Lough (*ELC*, 7 March 1876)



2

[Lough Family Letters (Miller's Collection). Donor: Eilish O'Mahony, 59 Anaverna Drive, Ard Easmuinn, Dundalk, Co. Louth]

Same as #1

[Letter Number 66]

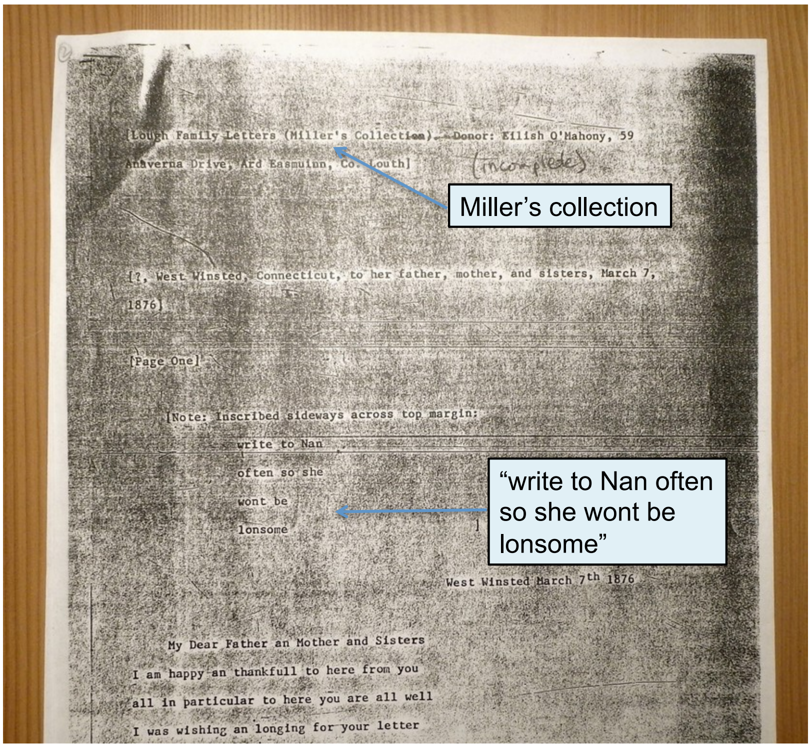
[Lizzie Lough, Winsted, Connecticut?, to her mother, ?]

Miller's collection – additional four pages

[Page One]

[Note: Inscribed sideways across top margin:
I must put
in another
word here
This is the first
fine day we
had this year
The snow is

Figure 5.3: Miller's transcription of the additional four pages belonging to the Elizabeth Lough letter (*ECL*, 7 March 1876)



[Lough Family Letters (Miller's Collection). Donor: Eilish O'Mahony, 59 Anaverna Drive, Ard Easmuinn, Co. Louth]

(incomplete)

[?, West Winsted, Connecticut, to her father, mother, and sisters, March 7, 1876]

[Page One]

[Note: Inscribed sideways across top margin:
write to Nan
often so she
wont be
lonsome
]

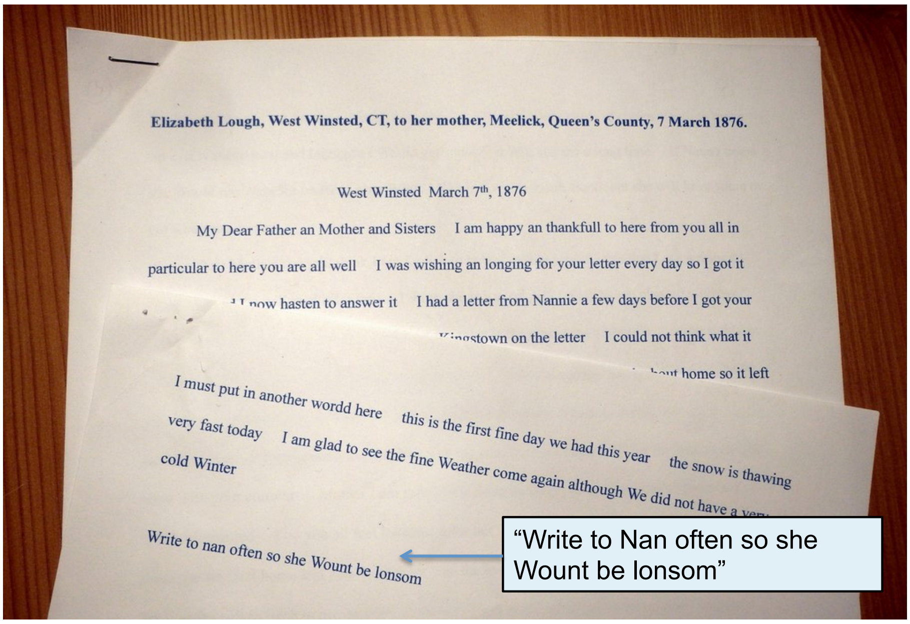
West Winsted March 7th 1876

My Dear Father an Mother and Sisters
I am happy an thankfull to here from you
all in particular to here you are all well
I was wishing an longing for your letter

Miller's collection

"write to Nan often so she wont be lonsome"

Figure 5.4: The complete Elizabeth Lough letter, transcribed by Miller (*ELC*, 7 March 1876)



Elizabeth Lough, West Winsted, CT, to her mother, Meelick, Queen's County, 7 March 1876.

West Winsted March 7th, 1876

My Dear Father an Mother and Sisters I am happy an thankfull to here from you all in particular to here you are all well I was wishing an longing for your letter every day so I got it

I now hasten to answer it I had a letter from Nannie a few days before I got your

Kingstown on the letter I could not think what it

home so it left

I must put in another wordd here this is the first fine day we had this year the snow is thawing very fast today I am glad to see the fine Weather come again although We did not have a very cold Winter

Write to nan often so she Wount be lonsom

"Write to Nan often so she Wount be lonsom"

Figure 5.5: Miller's RA's transcription of the complete Elizabeth Lough letter in MS Word (*ELC*, 7 March 1876)

Within this file, then, there are four iterations of the transcription; and within each of those iterations there are slight variations. For example, in the top left-hand margin Elizabeth writes the follow message: *write to nan / often so she / won't be lonesome*.[119] Table 5.1 summarises the differences – minutiae relating to spelling and capitalisation – in Schrier's, Miller's and Miller's RA's transcriptions. It should be stressed that this is not a criticism of the work carried out by Schrier and Miller. The fact that all of these transcriptions have been preserved and are available allows the transcription history of Elizabeth's letter to be fully documented. Furthermore, these relatively small differences in transcriptions would not have been crucial to Schrier's and Miller's research, which was mainly concerned with the content of the letter (*what* Elizabeth wrote about – her experiences and preoccupations). Nevertheless, spelling variations and the use of capitalisation are of special linguistic interest in some disciplines (historical and sociolinguistics, for instance), potentially revealing something about an author's schooling and social status.[120] Without these differences in transcription practices being documented or explained, and without access to the original manuscript, it is difficult to know which version of Elizabeth's letter is most reliable, which, in turn, may cast doubts on any research findings – depending, of course, on the nature of the research being carried out.[121]

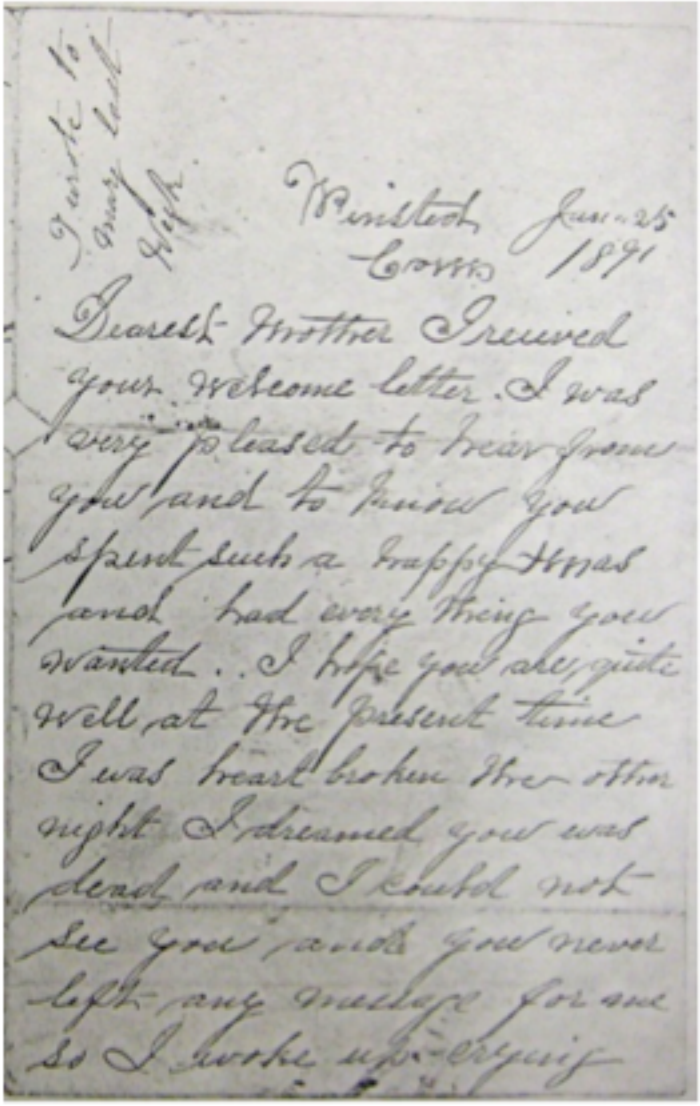
Schrier	**Miller**	**Miller's RA**
Write	write	Write
wount	wont	Wount
lonsom	lonsome	lonsome

Table 5.1: Differences in transcription practices[122]

[119] Note that the text here has been standardised for demonstration purposes; the forward slash represents a line break.
[120] See Fairman (2008; 2012).
[121] For more on the problems of digitally transcribing handwritten documents see Fairman (2015 *forth*).
[122] Note, in particular, differences in the use of capitalisation and spelling in Table 5.1.

In short, although not always possible, in an ideal situation, the researcher will have the original manuscript (document) together with a digital transcription of that manuscript (text) to hand, making it possible to constantly move between the two, identifying what has been lost, or gained, through the transcription process.

Looking at Figure 5.6 (another of the Lough letters – this time a letter by Julia Lough, the last sister to emigrate to America in 1884 at just 13 years of age) and Figure 5.7 (a digital transcription of Julia's letter, saved in Plain Text format), it is clear that there is little in common between the two in terms of structure and layout. In Figure 5.6, for instance, the address and date are right aligned at the top of the page and there is some text written vertically in the top left margin. The letter is handwritten and it is possible to see where lines begin and end, and where there is empty space (in the top right corner, for example). However, much of this information has been lost in the digital transcription. Reading the content of Figure 5.7 would reveal that this is a type of correspondence (there is a date, *Jan 25 1891*, and a salutation, *Dearest Mother*); however without information regarding the physical properties of the original document, it would be difficult to say for certain whether Figure 5.7 is one mode of correspondence or another (a handwritten letter, an email, or, for argument's sake, a transcription of a voice message for instance). The information that is lost when a document is transcribed and saved into a digital format needs to be captured and represented in some way. In other words, information that is implicit when looking at Figure 5.6 (the fact that it is a handwritten letter) needs to be made explicit when looking at Figure 5.7. This process is described as text encoding.



I wrote to
Mary last
Week

Winsted Jan..25
Conn 1891

Dearest Mother I recived
your welcome letter. I was
very pleased to hear from
you and to know you
spent such a happy Xmas
and had every thing you
wanted.. I hope you are quite
Well at the present time
I was heart broken the other
night I dreamed you was
dead and I could not
See you and you never
left any message for me
So I woke up crying

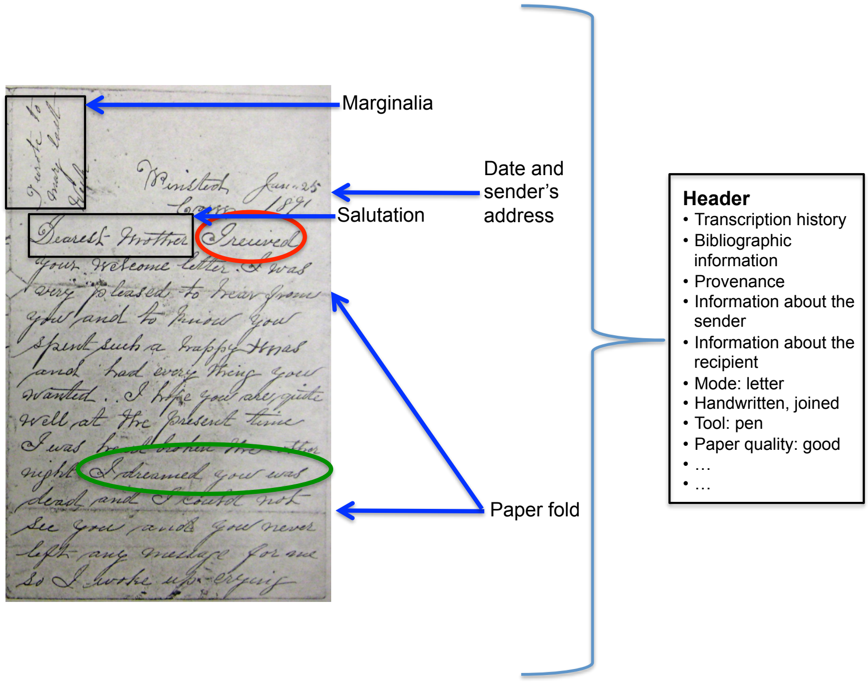
Figure 5.6: Original manuscript: page one of a letter by Julia Lough to her mother (*JLC*, 25 January 1891)

```
I wrote to Mary last Week Winsted Jan..25 Conn. 1891 Dearest Mother I
recived your welcome letter. I was very pleased to hear from you and
to know you spent such a happy Xmas and had every thing you wanted. I
hope you are quite Well at the present time I was heart broken the
other night I dreamed you was dead and I could not See you and you
never left any message for me so I woke up crying
```

Figure 5.7: Digital (Plain Text) version of the Julia Lough letter shown in Figure 5.6

The process of encoding is an intellectual activity, which involves thinking about which features of the document to represent, the relationship between those features and how those features should be named, described and categorised in a structured and formalised way. It is never a neutral process; the way a text is encoded will reveal something about what the encoder believes to be important or salient about the original document. The same text, then, can be encoded in many different ways, drawing out features that are most relevant to the encoder's own research interests, or the project aims. Encoding enriches the text and makes things visible. It has the potential of allowing a text to be looked at in new ways and from many disciplinary perspectives, offering different ways-in and providing different layers of interpretation.

Looking at Figure 5.8, for instance, it is possible to see a wealth of information that might be of interest to a broad range of scholars, with a broad range of research questions. A paper conservator or paper historian, for instance, may be interested in the quality of the paper or how the paper has been folded (vertically through the centre and horizontally one third from the top/bottom of the sheet), whereas a graphologist may wish to capture information about the handwriting style, whether there are multiple authors, whether the original letter is written in pen or in pencil and how the document is structured – perhaps distinguishing between writing that is contained in the margins and writing that is contained in the main body. Additionally, whereas linguists may be more interested in capturing information about the language of the document, such as spelling variants (*I recived* [*sic passim*], circled at the top), syntactic variants (*I dreamed you was dead*, circled at the bottom), omissions or repetitions, historians, on the other hand, may be more interested in any references to people, places and significant events. Finally, there may also be contextual information that is not explicitly stated within the document's content, such as that described in chapter one (Julia's age, when she emigrated, who she married, where she worked etc.), which scholars may wish to capture.



I wrote to Mary last week

Winsted Jan 25
Conn 1891

Dearest Mother I received your welcome letter. I was very pleased to hear from you and to know you spent such a happy Xmas and had every thing you wanted. I hope you are quite well at the present time I was heart broken the other night I dreamed you was dead and I could not see you and you never left any message for me So I woke up crying

Header
- Transcription history
- Bibliographic information
- Provenance
- Information about the sender
- Information about the recipient
- Mode: letter
- Handwritten, joined
- Tool: pen
- Paper quality: good
- …
- …

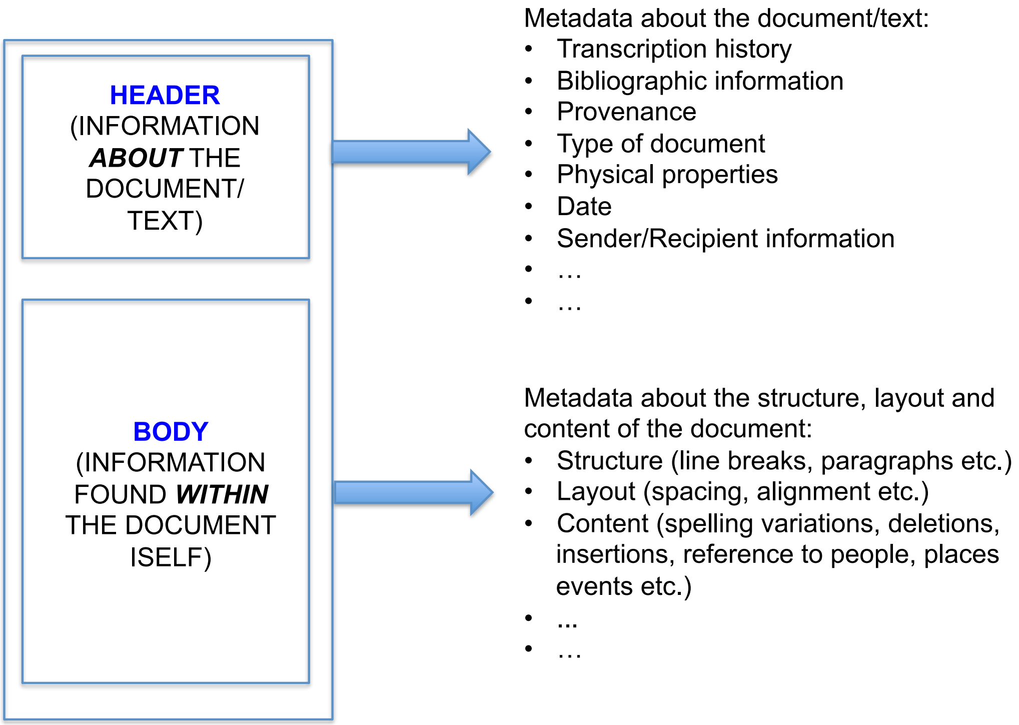
Figure 5.8: Possible features to encode (page one of a letter by Julia Lough to her mother (*JLC*, 25 January 1891)

All of this very rich and very useful information, relating to the letter's content and context, needs to be documented within the digital edition, as I will now discuss in more detail.

Text encoding involves adding metadata (in the form of tags) to the text. Metadata (sometimes described as data about data) provides additional information or knowledge about the content or context of the text. In the example that follows, the tags contain the following metadata: they state that 'Mary' is a person's name, represented using the tags <persName></persName>, and that 'Meelick' is a place name, represented using the tags <placeName></placeName>:

```
<persName>Mary</persName>  is still in <placeName>Meelick</placeName>
```

This metadata can be contained within the body of the text, providing information or knowledge regarding the structure (lines, paragraphs etc.), layout (spacing, alignment etc.), or content of the original document (spelling variations, deletions, insertions, or references to people, places or events, for instance). It can also be situated outside the body of the text in what is called the header,[123] providing information or knowledge about the document/text itself (see Figure 5.9).


HEADER
(INFORMATION ***ABOUT*** THE DOCUMENT/ TEXT)
BODY
(INFORMATION FOUND ***WITHIN*** THE DOCUMENT ISELF)

Metadata about the document/text:

- Transcription history
- Bibliographic information
- Provenance
- Type of document
- Physical properties
- Date
- Sender/Recipient information
- …
- …

Metadata about the structure, layout and content of the document:

- Structure (line breaks, paragraphs etc.)
- Layout (spacing, alignment etc.)
- Content (spelling variations, deletions, insertions, reference to people, places events etc.)
- ...
- …

Figure 5.9: Two levels of encoding

This chapter will focus on the header information as this poses fewer barriers in terms of accessibility and intellectual property issues; that is to say,

[123] As will be explained later in the chapter, in TEI markup language the body is marked using the tag <body></body> and the header is represented using the tag <header></header>.

projects are often more willing to give access to information about a letter collection than they are to give access to the collection itself. For reasons of manageability, I decided to subdivide the header into three layers (highlighted in the box in Figure 5.1 at the beginning of this chapter (p. 173)). First, ‘Document/Text’ captures information about the document and/or text itself, such as the transcription history (the number of iterations, their format and the person responsible, for instance) and transcription practices (details about the methodological decisions made as regards representing capital letters and spelling variations, for instance), as well as bibliographic information and details about the document’s provenance. Second, ‘Personography’ captures information about whom the letter is from (the sender) and whom the letter is to (the recipient). Third, ‘Placeography’ captures information about the location of the sender and the location of the recipient. Some of the header information can be gleaned from the letter content. Details of the sender and recipient, for instance, may well be found in the salutation and sign-off contained within the body of the letter. However, as discussed previously, the researcher often has a lot of additional information regarding the participants involved in the act of communication, which is not explicitly stated in the letter, but which needs capturing nonetheless (information such as date of birth, marriage, death, number of children, occupation/s and so on).

I have chosen to use the TEI (Text Encoding Initiative) Guidelines (2008) to carry out the encoding process. There are many ways of encoding; however, the TEI is the de facto standard for encoding digitised texts in the humanities, providing a markup language and guidelines. TEI is multilingual (so it can be used for letter collections in different languages), it is open-ended (so it can accommodate new things), and it is flexible, subjective and interpretative. And, finally, TEI represents consensus within a research community. Specifically, the TEI,

> ...make recommendations about suitable ways of representing those features of textual resources which need to be identified explicitly in order to facilitate processing by computer programs, suggesting a set of markers (or tags) which may be inserted in the electronic representation of the text, in order to mark the text structure and other features of interest' (TEI Consortium 2008, p. xxiii).

Although a relatively labour intensive process, once encoded, computer programs can then be used to search and visualise texts in various ways. In the case of emigrant letters, this might involve pulling out all references to place names and presenting them on a map, for instance, or visualising letter writing networks. To put it simply, encoding makes explicit to a computer what is implicit to a person and most computer programs will '…depend on the presence of such explicit markers for their functionality, since without them a digitized text appears to be nothing but a sequence of undifferentiated bits' (TEI Consortium 2008, p. xxiii).

TEI markup language

As previously mentioned, the TEI provides a markup language and guidelines for encoding texts in the humanities. The TEI describe the term 'markup language' as being a 'set of … conventions used together for encoding texts'. A markup language 'must specify [1] how markup is to be distinguished from text, [2] what markup is allowed, [3] what markup is required and [4] what the markup means' (TEI Consortium 2008, p. xxxi).

TEI markup language is an application of XML (eXtensible Markup Language). There are various applications of XML, such as the MEI (Music Encoding Initiative) markup language[124] and MathML (Mathematical Markup Language),[125] but whereas the TEI is concerned with digitally representing texts in the humanities (poems, plays, historical manuscripts, dictionaries and so on), the MEI is concerned with the digital representation of music notation documents and MathML is concerned with the digital representation of mathematical notations.

As an application of XML, TEI markup language must do certain things for it to be XML compliant – that is, it must obey certain rules for it to be considered well-formed. Those rules (as explained in the *TEI Guidelines*) are summarised below:

1. A tag (in the form of angle brackets <>) must explicitly mark the start and end of each *element*. Elements represent the structural components of a text, such as the body, paragraphs and line breaks. Elements of one type (say, a paragraph) may be embedded within elements of another type (the body, for instance). In the example on the following page,[126] there are two paragraphs, where <p> represents the start of a paragraph and </p> (with forward slash) represents the

[124] *Music Encoding Initiative* Available from: http://en.wikipedia.org/wiki/Music_Encoding_Initiative [Accessed 1 March 2015].

[125] *MathML* Available from: http://en.wikipedia.org/wiki/MathML [Accessed 1 March 2015].

[126] The markup being used here is just for demonstration purposes (that is to say, it does not reflect the structure and layout of the original manuscript).

end of a paragraph. These paragraphs are embedded within the body (<body></body>) and each paragraph contains several line breaks (<lb/>).[127]

```
<body>
    <p>
        <lb/>Dearest Mother I recived
        <lb/>your welcome letter. I was
        <lb/>very pleased to hear from
        <lb/>you and to know you
        <lb/>spent such a happy Xmas
        <lb/>and had every thing you
        <lb/>wanted.
    </p>
    <p>
        <lb/>I hope you are quite
        <lb/>well at the present time
    </p>
</body>
```

2. There must be a single element enclosing the whole document: this is known as the root element. In the example above, <body></body> represents the root element – all other elements (paragraphs <p></p> and line breaks <lb/>) are contained within this root element.
3. Each element apart from the root element must be completely contained by another element; elements cannot partially overlap one another. The following example, for instance, would not be XML compatible as the second paragraph has no closing tag and is therefore not completely contained by the root element (in this case, <body></body>).

[127] Note that line breaks operate slightly differently – these are known as 'empty elements'. They do not have an opening and closing tag. Instead the tag <lb/> is used at the point in the text where a new line starts (for more information about <lb/> see: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-lb.html).

```
<body>
    <p>
        <lb/>Dearest Mother I recived
        <lb/>your welcome letter. I was
        <lb/>very pleased to hear from
        <lb/>you and to know you
        <lb/>spent such a happy Xmas
        <lb/>and had every thing you
        <lb/>wanted.
    </p>
    <p>
        <lb/>I hope you are quite
        <lb/>well at the present time
</body>
```

4. In addition to elements, there are *attributes*. Each element can posses one or more attributes. The TEI describe attributes as ‘information that is in some sense descriptive of a specific element occurrence but not regarded as part of its content’ (TEI Consortium 2008, p. xlii). Perhaps another way of understanding elements and attributes is to think about them as nouns and adjectives. Nouns are naming devices, while adjectives describe, classify and further categorise that noun. So in the example below, the <title> element tells us that what follows is a title and the @level attribute gives more information about what type of title it is (in this case it is a series title as indicated by the "s" – note that the attribute value (in this case "s") must be quoted). Additionally, it has become common convention to use the XPath notation for attribute names, by prefixing them with the @ sign (that is to say, the ‘@’ symbol says that what follows is an attribute).

```
<title level=“s”>Irish Emigrant Letters</title>
```

5. Other key characteristics of well-formed XML are the case-sensitivity of tag names and the fact that special reserved characters ('<' and '&') must be escaped with entity references.[128]

Provided the above rules are adhered to, a text will be considered well-formed XML.

From document analysis to markup

The first stage of the encoding process involved carrying out a detailed analysis of the document/s to be digitised and marked-up, identifying features that are important in and across the different letter collections I have been working with (including the Lough letters as well as the *MALE Ref.* and *FEMALE Ref.* corpora discussed in chapters two and three).[129] Fortunately, as part of an AHRC funded research networking project entitled *Digitising experiences of migration: the development of interconnected letter collections* (DEM),[130] I had the opportunity to repeat this process with a group of scholars from a range of disciplinary backgrounds, including historians, migration studies experts, archivists, socio-linguists, corpus linguists and digital humanists.[131] When analysing a selection of

[128] Certain characters cannot be used within XML because they have special meanings. These include the ampersand (&), "double quotes", 'single quotes' and <angle brackets>. Many of the Lough letters contain, for example, the ampersand in the main body of the text, so these characters need to be 'escaped' with a predefined entity reference as follows: **&amp;**. For more on entity references see: *TEI by Example* (2004) Available from: http://www.teibyexample.org/modules/TBED00v00.htm?target=xmlgroundrules [Accessed 1 August 2015].

[129] For this stage of the process I referred to several collections within Professor Miller's archive, including the Lough collection.

[130] AHRC Project reference: AH/K006231/1.

[131] Full details of the research network partners, together with the letter collections they are involved with, can be found on the project blog: http://www.lettersofmigration.blogspot.com.

the Lough letters, the research network partners were asked to consider four key questions:

1) What do different researchers use correspondence collections for?
2) What features of the letters are considered important across the disciplines?
3) What is unique, special or different about emigrant correspondence?
4) In what ways would you like to be able to search and visualise emigrant letters?

Through discussing these questions during one of the DEM project workshops it was possible to identify where there were commonalities and differences across the disciplines (in terms of how emigrant letters are being used, what sort of information is viewed as important and why, and how that information is described and labelled) and where reduplication was taking place (instances where scholars from different disciplines appeared to be doing very similar work with regard to the digitsation, documentation and analysis of correspondence collections – in some cases even working on the same letters).

Table 2 summarises the information which the research network partners wanted to capture in relation to 'Document/Text', 'Personography' and 'Placeography', within the TEI header. Underlined are the features that were felt to be especially important as regards emigrant letters, namely provenance (where the letter originates from; how it came into being), relationships (letter writing networks) and locations (the movement of migrants over time).[132]

[132] I would like to formally thank the project partners for their involvement in, and contributions to, the DEM project. Full details of the workshops can be found on the project blog: http://www.lettersofmigration.blogspot.fr. Specifically, this chapter builds on discussions from workshop one in which participants discussed what features of emigrant letters would be 'desirable' and/or 'essential' to capture in a system of markup. The outcomes of that discussion

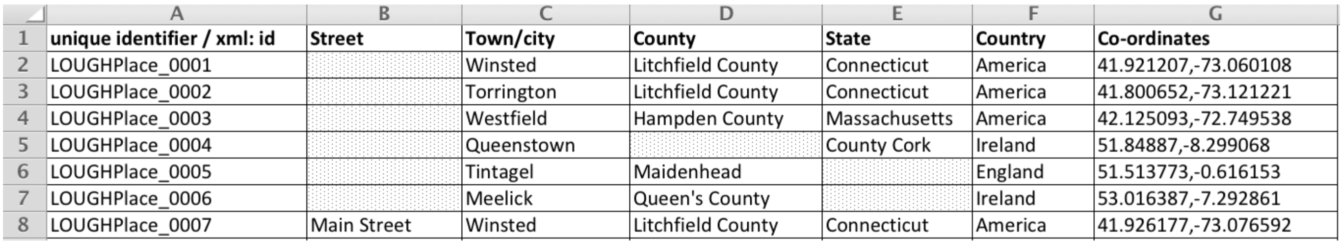
There are three things to note when looking at Table 5.2. First, only those features of emigrant letters that project partners agreed to be significant across the disciplines are captured. In other words, the list is not exhaustive and different emigrant letter projects will have different requirements and may wish to capture different information. Second, it is rare that all of the information listed in Table 5.2 will be available for any given letter. Quite often researchers are working with fragmented or damaged manuscripts and it is not possible to establish even very basic information such as the sender or the recipient. Third, if this information is captured in three simple spreadsheets (one spreadsheet containing information about the document/text, another containing personography information and a final spreadsheet containing placeography information (see Figure 5.10), it is much easier to manage the metadata and to notice where there are gaps or duplications. Additionally, it is relatively straightforward – from a programming perspective – to convert the metadata contained within these three spreadsheets into TEI compliant XML.[133]

are summarised in Table 5.2. In this chapter, I propose a TEI markup template which would allow that information to be captured in a formalised and structured way, which, if applied across letter collections, would avoid issues relating to differing terminologies across disciplines and the reduplication of work.

[133] Note that a programmer would be needed to write the necessary script.

Document/text	**Personography information** *Sender/Recipient*	**Placeography information** *Sender/Recipient*
• Transcription history • Bibliographic information • Provenance statement • Date of letter	• First name • Surname • Maiden name • Married name • Nickname/s • Date of birth • Date of death • Date of marriage • Sex • Occupation/s • Social class • Education • Faith • Relationships • Date emigrated • Residences • Additional information	• Street • Village, Town, City • Region • Country • GIS coordinates (latitude/longitude) • Additional information

Table 5.2: Three layers of metadata for the TEI header



	A	B	C	D	E	F	G
1	**unique identifier / xml: id**	**Street**	**Town/city**	**County**	**State**	**Country**	**Co-ordinates**
2	LOUGHPlace_0001		Winsted	Litchfield County	Connecticut	America	41.921207,-73.060108
3	LOUGHPlace_0002		Torrington	Litchfield County	Connecticut	America	41.800652,-73.121221
4	LOUGHPlace_0003		Westfield	Hampden County	Massachusetts	America	42.125093,-72.749538
5	LOUGHPlace_0004		Queenstown		County Cork	Ireland	51.84887,-8.299068
6	LOUGHPlace_0005		Tintagel	Maidenhead		England	51.513773,-0.616153
7	LOUGHPlace_0006		Meelick	Queen's County		Ireland	53.016387,-7.292861
8	LOUGHPlace_0007	Main Street	Winsted	Litchfield County	Connecticut	America	41.926177,-73.076592

Figure 5.10: Placeography information relating to the *LOUGH Corpus*, captured in an Excel spreadsheet

The next stage in the encoding process involved deciding on how best to describe and organise the information listed in Table 5.2 within the TEI header, thus formalising and standardising the metadata, thereby allowing letter collections to potentially interconnect. Figure 5.11 shows the header markup for the Elizabeth Lough letter mentioned at the beginning of this chapter. I will now discuss my proposed markup template in more detail.

```
<teiHeader xmlns="http://www.tei-c.org/ns/1.0">
    <fileDesc>
        <titleStmt>
            <title level="s">Irish Emigrant Letters</title>
            <title level="a">Elizabeth Lough to her parents and sisters<lb/>Winsted, 7 March 1876</title>
            <author key="LOUGHPers_0001">Elizabeth Lough</author>
            <editor>Emma Moreton</editor>
        </titleStmt>
        <editionStmt>
            <edition>Digitising Experiences of Migration (DEM)</edition>
            <respStmt>
                <resp>Typed transcription of original manuscript</resp>
                <name>Professor Arnold Schrier</name>
                <name>Professor Kerby Miller</name>
            </respStmt>
            <respStmt>
                <resp>MS Word version of Miller’s typed transcription</resp>
                <name>Miller’s RA</name>
            </respStmt>
            <respStmt>
                <resp>Digital version based on Miller’s and Schrier’s typed transcriptions</resp>
                <name>Emma Moreton</name>
            </respStmt>
            <respStmt>
                <resp>Converted to XML format and markup added</resp>
                <name>Emma Moreton</name>
            </respStmt>
        </editionStmt>
        <publicationStmt>
            <publisher>Coventry University</publisher>
            <availability status="restricted">
                <p>Available under a CC-BY license</p>
            </availability>
        </publicationStmt>
        <sourceDesc>
            <msDesc>
                <msIdentifier>
                    <repository>Professor Kerby Miller, History Department, University of Missouri</repository>
                    <collection>Lough Family Letters</collection>
                    <idno>LOUGH_001</idno>
                </msIdentifier>
                <history>
                    <p>There are 99 letters in the Lough collection. In the early 1950s, a few of the Lough
                    letters were donated to Arnold Schrier (Professor Emeritus, University of Cincinnati).
                    In the 1970s and 1980s, the rest of the Lough letters were donated to Kerby Miller
                    (Curators' Professor, University of Missouri) by the O'Mahonys and by Edward Dunne
                    and Mrs Kate Tynan of Portlaoise, County Laois. Both Miller and Schrier made
                    transcriptions of the letters and returned the original manuscripts to the donors. The
                    collection contains photocopies of the original manuscripts together with the typed
                    transcriptions.</p>
                </history>
            </msDesc>
        </sourceDesc>
    </fileDesc>
    <profileDesc>
        <ct:correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-force-correspDesc">
            <ct:participant role="sender"
                  <persName key="LOUGH_Pers0001">Elizabeth Lough</persName>
                  <placeName key="LOUGH_Place0001">Winsted, Connecticut</placeName>
                  <date when="1876-03-07"/>
            </ct:participant>
            <ct:participant role="recipient"
                <persName key="LOUGH_Pers0007">Elizabeth McDonald Lough</persName>
                <persName key="LOUGH_Pers0008">James Lough</persName>
                <persName key="LOUGH_Pers0002">Alice Lough</persName>
                <persName key="LOUGH_Pers0003">Anne Lough</persName>
                <persName key="LOUGH_Pers0004">Julia Lough</persName>
                <persName key="LOUGH_Pers0005">Mary Lough</persName>
                <placeName key="LOUGH_Place0006">Meelick, Queen’s County</placeName>
            </ct:participant>
            <langUsage>
                <language> ident="en">English</language>
            </langUsage>
        </ct:correspDesc>
    </profileDesc>
</teiHeader>
```

Figure 5.11: Example of the header markup using Elizabeth’s 7 March 1876 letter as an example

There are four possible sections within the TEI header: <fileDesc> (file description), <encodingDesc> (encoding description), <profileDesc> (profile description) and <revisionDesc> (revision description). The file description <fileDesc> is a mandatory section within all TEI headers, 'containing a full bibliographic description of the computer file itself, from which a user of the text could derive a proper bibliographic citation…' (TEI Consortium 2008, p. 17); the remaining three sections (<encodingDesc>, <profileDesc> and <revisionDesc>) are optional. However, for correspondence projects, it is necessary to include the <profileDesc> as this is where information about the correspondence itself, namely information about the sender and recipient (name and location) and the date of the letter, is captured.

To summarise, then, for (emigrant) letter projects, the header should include, as a minimum, the following two sections: <fileDesc> and <profileDesc> (see Figure 5.12). Other sections can, of course, be added and this will depend on factors such as the project requirements, time and budget; however, the aim of this chapter is to provide a basic, skeleton markup for encoding emigrant letters, which can be applied, and built on, across letter collections.

I will now look at <fileDesc> (where information about the document/text is captured) and <profileDesc> (where personography and placeography information is captured) in turn.

```
<teiHeader>
    <fileDesc></fileDesc>
    <profileDesc></profileDesc>
</teiHeader>
```

Figure 5.12: Recommended sections within the TEI header for emigrant letter projects

File description <fileDesc>

The file description <fileDesc> (see Figure 5.13) contains three mandatory elements <titleStmt> (title statement), <publicationStmt> (publication statement) and <sourceDesc> (source description) and one optional element <editionStmt> (edition statement).

```
<teiHeader>
    <fileDesc>
        <titleStmt></titleStmt>
        <editionStmt></editionStmt>
        <publicationStmt></publicationStmt>
        <sourceDesc><sourceDesc>
    </fileDesc>
</teiHeader>
```

Figure 5.13: <fileDesc> (file description)

Title statement <titleStmt>

Within <titleStmt> (see Figure 5.14) there are three elements: <title>, <author> and <editor>. The <title> element has been subdivided into two levels using the *@level* attribute: @level="s" is used to describe the series or collection to which the text belongs; @level="a" is used to describe the analytic item (the text itself).

The letters referred to in this chapter are by Irish emigrants and are therefore categorised as belonging to the ‘Irish Emigrant Letters’ collection, rather than, say, the ‘Portuguese Emigrant Letters’ collection. The analytic title provides more detail about the text itself, including the author’s name, the recipients of the letter, and the date: ‘Elizabeth Lough to her parents and sisters’ and on a separate line (marked by <lb/> to indicate a line break) ‘Winsted, 7 March 1876’. Obviously, the titles will vary depending on the project; however, where possible, titles should be consistent and provide a clear, meaningful description of the text in question (rather than being, for instance, a reference number).

```
<titleStmt>
    <title level="s">Irish Emigrant Letters</title>
    <title level="a"> Elizabeth Lough to her parents and sisters<lb/>Winsted, 7 March
    1876</title>
    <author key="LOUGHPers_0001"/>Elizabeth Lough</author>
    <editor>Emma Moreton</editor>
</titleStmt>
```

Figure 5.14: <titleStmt> (title statement)

The <author> element provides the name/s of the author/s responsible for the content of the manuscript. The <author> is typically the same as the "sender" within the <ct:correspDesc> element, as will be discussed later; however, there may be times when the author and sender are different. It is possible, albeit unusual, perhaps, for a newspaper clipping to be posted without a covering letter. In this instance the author of the newspaper clipping would not be the same as the sender (the person responsible for posting the newspaper clipping).

Sometimes there is only limited information available as regards the author (there may, for example, just be a first name – if that) and sometimes there is a lot of information (information to do with the author’s family history, their occupation/s, their education and so on). Rather than capturing this information within the header itself, which can make the header somewhat cluttered and less manageable, it is possible to use a pointing mechanism which directs the user to a separate personography file, containing all information about that person. I will write in more detail about the personography file later; at this point, however, it is worth focusing briefly on the options regarding pointing mechanisms. The *TEI Guidelines* offer two possibilities in this regard: the @ref attribute and the @key attribute. Full details and explanations of these attributes can be found in the *TEI Guidelines*; however, in summary, the @ref attribute 'provides an explicit means of locating a full definition for the entity being named by means of a URI [uniform resource identifier]' (TEI Consortium 2008, p. 402) – in other words, it points to (an identified fragment in) an external file with its own URI; the @key attribute, in contrast, points to its 'own local database system containing canonical information about persons and places, each entry in which is accessed by means of some system-specific identifier constructed in a project-specific way' (TEI Consortium 2008, p. 392) – in other words, it points to an internal database system of some kind.

The *TEI Guidelines* recommend that as 'no particular syntax is proposed for the values of the key attribute, since its form will depend entirely on practice within a given project' it is 'not recommended in data interchange, since there is no way of ensuring that the values used by one project are distinct from those used by another'.[134] In terms of interoperability, data sharing and accessibility of resources, then, the @ref attribute is preferred. In those cases where copyright, ethical and intellectual property issues prevent a collection (or information relating to that collection) from being made freely available there are two possibilities: 1) use the @key attribute (as is the case for the Lough letters referred to in this chapter) or 2) specify a relative URI to project-internal personography files using the @ref attribute.[135]

The final element within <titleStmt> is the <editor> element. Here, information about the person ultimately responsible for creating the digital text

[134] *TEI Guidelines*. Available from: http://tei.oucs.ox.ac.uk/release/doc/tei-p5-doc/en/html/ref-att.canonical.html [Accessed 7 May 2015].

[135] For more information on URIs see the W3C recommendations at http://www.w3.org/TR/REC-html40/intro/intro.html#h-2.1.3 [Accessed 7 May 2015].

and overseeing the encoding process can be captured. More details regarding the transcription process (who was involved and at what stage) can be provided in the <editionStmt>, which is discussed in the following section.

Edition statement <editionStmt>

The <editionStmt> is optional; however, it is recommended that emigrant letter projects include this as it allows information about the transcription history to be captured in a formalised way. As previously discussed, often there are different iterations, or editions of the same manuscript. In the example given at the beginning of this chapter (the letter by Elizabeth Lough), there were five people involved in the transcription process (see Figure 5.15). Schrier and Miller both produced typed transcriptions. Miller's RA produced a digital version in MS Word, based on Miller's transcriptions, and I produced a digital version in Plain Text format based on Schrier's and Miller's transcriptions. I then produced a marked-up version of the text in XML, which I passed to the TEI Correspondence SIG for comment.[136]

[136] I would like to thank, in particular, Peter Stadler, Universität Paderborn – one of the convenors of the TEI SIG: Correspondence – for his very useful feedback.

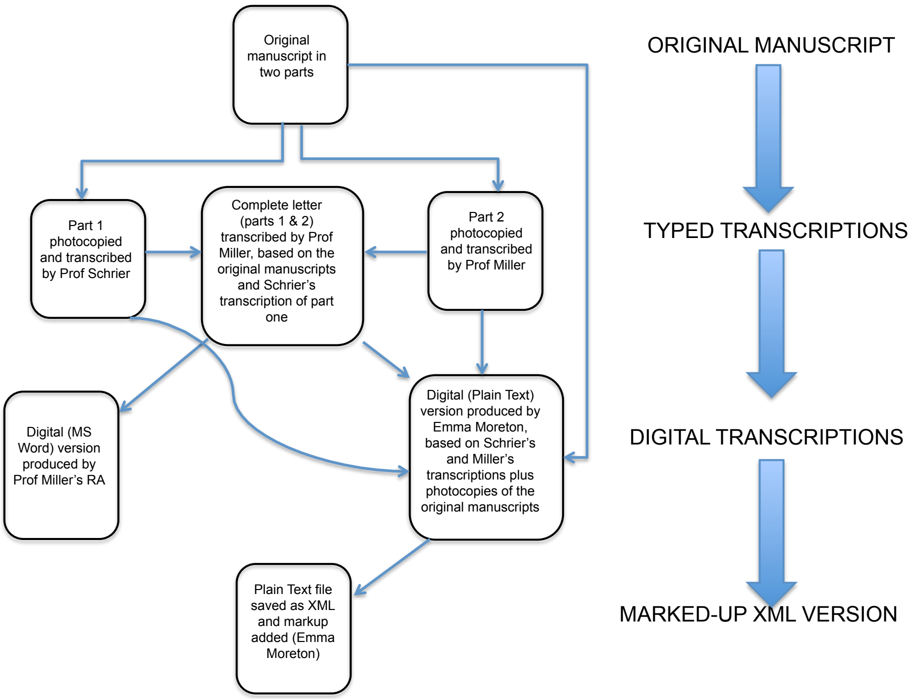

Original manuscript in two parts
Part 1 photocopied and transcribed by Prof Schrier
Complete letter (parts 1 & 2) transcribed by Prof Miller, based on the original manuscripts and Schrier's transcription of part one
Part 2 photocopied and transcribed by Prof Miller
Digital (MS Word) version produced by Prof Miller's RA
Digital (Plain Text) version produced by Emma Moreton, based on Schrier's and Miller's transcriptions plus photocopies of the original manuscripts
Plain Text file saved as XML and markup added (Emma Moreton)
ORIGINAL MANUSCRIPT
TYPED TRANSCRIPTIONS
DIGITAL TRANSCRIPTIONS
MARKED-UP XML VERSION

Figure 5.15: Transcription history of Elizabeth Lough letter (*ELC*, 7 March 1876)

All of this information can be captured within <editionStmt>. Looking at Figure 5.16, the <edition> element provides details regarding which edition of the text this is (in this case, the digital transcription was produced as part of the DEM project, mentioned previously). Listed underneath, the <respStmt> (statement of responsibility) is used to outline the process which led to the creation of the digital text. The <resp> (responsibility) element details the action ('produced typed transcription', 'produced digital transcription', 'converted to XML', for example) and the <name> element details the person/s responsible for that action.

```
<editionStmt>
    <edition>Digitising Experiences of Migration (DEM)</edition>
    <respStmt>
        <resp>Typed transcription of original manuscript</resp>
        <name>Professor Arnold Schrier</name>
        <name>Professor Kerby Miller</name>
    </respStmt>
    <respStmt>
        <resp>MS Word version of Miller's typed transcription</resp>
        <name>Miller's RA</name>
    </respStmt>
    <respStmt>
        <resp>Digital version based on Miller's and Schrier's typed transcriptions</resp>
            <p>Original spelling and typography retained</p>
        <name>Emma Moreton</name>
    </respStmt>
    <respStmt>
        <resp>Converted to XML format and markup added</resp>
        <name>Emma Moreton</name>
    </respStmt>
</editionStmt>
```

Figure 5.16: <editionStmt> (edition statement)

Publication statement <publicationStmt>

The element <publicationStmt> is mandatory (see Figure 5.17). It contains 'information concerning the publication or distribution of an electronic or other text' (TEI Consortium 2008, p. 26). For example, as is often the case with large-scale letter projects, the original manuscripts might be held in archives around the world, but the digital transcriptions (the texts) might be published, or held on a server, elsewhere. The <publicationStmt>, then, gives information about the publication and distribution of the digital text. In this case, the <publisher> of the digital, marked-up text (the Elizabeth Lough letter) is Coventry University (my employer). Details of where the original manuscript is archived can be captured under <sourceDesc>, as will be discussed later. The <availability> element provides information about the accessibility of the text while the @status attribute can be used to show whether the digital text is open access or restricted, for instance.[137]

```
<publicationStmt>
    <publisher>Coventry University</publisher>
    <availability status="restricted">
        <p>Available under a CC-BY license</p>
    </availability>
</publicationStmt>
```

Figure 5.17: <publicationStmt> (publication statement)

Source description <sourceDesc>

```
<sourceDesc>
    <msDesc>
        <msIdentifier>
            <repository>Professor Kerby Miller, History Department, University of
            Missouri</repository>
            <collection>Lough Family Letters</collection>
            <idno>LOUGH_001</idno>
        </msIdentifier>
        <history>
            <p>There are 99 letters in the Lough collection. In the early 1950s, a few of the
            Lough letters were donated to Arnold Schrier (Professor Emeritus, University of
            Cincinnati). In the 1970s and 1980s, the rest of the Lough letters were donated
            to Kerby Miller (Curators' Professor, University of Missouri) by the O'Mahonys
            and by Edward Dunne and Mrs Kate Tynan of Portlaoise, County Laois. Both
            Miller and Schrier made transcriptions of the letters and returned the original
            manuscripts to the donors. The collection contains photocopies of the original
            manuscripts together with the typed transcriptions.</p>
        </history>
    </msDesc>
</sourceDesc>
```

Figure 5.18: <sourceDesc> (source description)

The final element within the <fileDesc> is the <sourceDesc> (see Figure 5.18). This is a mandatory element. It ‘supplies a description of the source text/s from which an electronic text was derived or generated’ (TEI Consortium 2008, p. 30).

[137] The CC BY license ‘lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials’ (see: http://creativecommons.org/licenses/).

Within <sourceDesc> is the <msDesc> (manuscript description) element. This element ‘contains a description of a single identifiable manuscript’ (TEI Consortium 2008, p. 31). Here, it is possible to capture information about the original manuscript such as the size and quality of the paper, watermarks, whether the letter included attachments and so on. However, as an absolute minimum, within <msDesc> there should be information which allows the user to locate the original manuscript. This information can be captured using the element <msIdentifier> (manuscript identifier). Within <msIdentifier> the elements <repository> (‘the name of a repository within which manuscripts are stored, possibly forming part of an institution’ (TEI Consortium 2008, p. 302)), <collection> (‘the name of a collection of manuscripts, not necessarily located within a single repository’ (TEI Consortium 2008, p.303)) and <idno> (‘any standard or non-standard number used to identify a bibliographic item’ (TEI Consortium, 2008, p. 303)) are used to record the precise details of the manuscript’s location. So, in Figure 5.18, the markup shows that this manuscript is part of a repository belonging to Professor Miller in the History Department at the University of Missouri. The manuscript is part of the ‘Lough Family Letters’ and its unique reference is ‘LOUGH_001’ (i.e. it is the first letter in the series). This information will allow users of the text to locate the original source should they wish to check or build on any aspect of the digital transcription or its markup.

It is often the case with emigrant correspondence that there is a great deal of family legend surrounding a letter and the people mentioned therein. All of this very valuable information, gathered over time from conversations with and between family members, can be captured within <sourceDesc>. As previously mentioned, provenance was deemed to be a particularly important aspect of emigrant letter collections. Capturing as much information as possible regarding the history of the text – how the text came into being, as well as information about the donors – will help future users of the resource to contextualise the letter content.

The *TEI Guidelines* define a specific child element of <msDesc> for this kind of provenance information, namely <history>, containing either a free prose description in <p> elements, or a structured description of different events concerning the origin, provenance, or acquisition of a manuscript, each in dedicated <origin>, <provenance> and <acquisition> elements. These elements are also members of the att.datable class, allowing for formal dating of the events they describe (TEI Consortium 2008, pp. 324-325). Technically, there is no reason why prose in the header cannot be searched as easily as any other text in an XML document; capturing provenance information within <p> elements should, therefore, pose no issues in terms of searchability. However, projects may prefer more nuanced and predictable search locations for structured queries relating to a letter's provenance, and this is something that TEI offers in <history>.

Profile description <profileDesc>

I will now look at the <profileDesc> section of the header (see Figure 5.19). The <profileDesc> 'provides a detailed description of non-bibliographic aspects of a text, specifically the languages and sublanguages used, the situation in which it was produced, the participants and their setting' (TEI Consortium 2008, p. 45).

As previously mentioned, the *TEI Guidelines* provide a standard for modelling different types of text in the humanities; however, until recently (April 2015), a separate correspondence module was lacking (Seifert *et al*. 2014, p. 2; Illetschko 2014, p. 1). In 2008 the TEI SIG: Correspondence was established to look specifically at issues to do with the markup of letter collections. The convenors of the SIG are Peter Stadler (Universität Paderborn), Marcel Illetschko (Austrian National Library, Vienna) and Sabine Seifert (Humboldt University, Berlin). For more information about the SIG, and to follow their progress, see the Correspondence SIG wiki at http://wiki.tei-c.org/index.php/SIG. And for example XML files, showing applications of the correspondence (correspDesc) module see https://github.com/TEI-Correspondence-SIG/correspDesc. At the time of writing this chapter, the TEI Correspondence SIG were proposing a special purpose container element, <ct:correspDesc>, which allows features of letters such as sender and recipient to be represented in a standardised way. This special purpose element is embedded under <profileDesc> within the TEI header. [138]

[138] Since writing this chapter the <correspDesc> proposal has been updated and now centres around <correspAction> and <correspContext> elements inside <correspDesc>, for describing details about the sending and reception of a letter as well as the context in which the letter occurs. However, compatibility between the markup template outlined in this chapter and any further evolutions of the <correspDesc> proposal should not be a problem. As mentioned previously, if header information relating to 'Document/Text', 'Personography' and 'Placeography' is captured in spreadsheets it is relatively straight forward for this information to be mapped onto the <correspDesc> module which was integrated into the *TEI Guidelines* in April 2015 (see: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-correspDesc.html). Appendix A demonstrates how the markup detailed in this chapter can be mapped onto the TEI's module for correspondence.

```
<profileDesc>
    <ct:correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-
    force-correspDesc">
        <ct:participant role="sender"
            <persName key="LOUGH_Pers0001">Elizabeth Lough</persName>
            <placeName key="LOUGH_Place0001">Winsted, Connecticut</placeName>
            <date when="1876-03-07"/>
        </ct:participant>
        <ct:participant role="recipient"
              <persName key="LOUGH_Pers0007">Elizabeth McDonald Lough</persName>
              <persName key="LOUGH_Pers0008">James Lough</persName>
              <persName key="LOUGH_Pers0002"/>Alice Lough</persName>
              <persName key="LOUGH_Pers0003"/>Anne Lough</persName>
              <persName key="LOUGH_Pers0004"/>Julia Lough</persName>
              <persName key="LOUGH_Pers0005"/>Mary Lough</persName>
              <placeName key="LOUGH_Place0006"/>Meelick, Queen's County</placeName>
        </ct:participant>
    </ct:correspDesc>
    <langUsage>
        <language> ident="en">English</language>
    </langUsage>
</profileDesc>
```

Figure 5.19: <profileDesc> (profile description)

Within <ct:correspDesc> the container element <ct:participant> captures information about the participants involved in the act of communication. The @role attribute allows those participants to be categorised as either sender or recipient, <ct:participant role="sender"> and <ct:participant role="recipient">, while the <persName> element lists the names of the sender and recipient/s. As previously discussed, with reference to the <author> element, information about the sender and/or recipient can be organised and managed in separate XML files; these are described as personography files. Personography files are effectively the same as authority files – a term used by archivists and librarians to describe bibliographic master files – and can record information such as the participant's date of birth/death, first name, surname, maiden name, nicknames, sex, occupation/s etc. Each personography file contains a <person> element with a unique identifier.[139] This identifier is used to create an association between the reference in the header (the value assigned to the relevant @key attribute) and the corresponding personography file.

Exactly the same process is used for managing information relating to the sender's/recipient's location. Details about the different locations (information such as the street name, town, region, country, as well as geographical coordinates) can be organised and managed in separate XML placeography files. Each placeography file contains a <place> element with a unique identifier which corresponds with the value assigned to the relevant @key attribute in the header. Having separate personography and placeography files (or authority files) for each person and location makes it much easier to manage changes to the metadata at a later date. (It is easier to change one master document – the personography or placeography file – than it is to change hundreds of documents.)

Within <ct:participant role="sender">, in addition to the personography and placeography information, there is a <date> element, which captures details of when the letter was dated. There are different ways to capture this information within the header. If the letter contains a specific date, then the year, month and day can be represented in the markup. However, quite often an exact date is missing and it is up to the researcher to make an educated guess as to when the letter was written. In such instances, the @notBefore and @notAfter attributes can be used to place the letter within an approximate timeframe: <date notBefore="1800" notAfter="1899"/>.

Finally, within the <profileDesc> element of the header it is also possible to capture information about the language of the letter using the <langUsage>

[139] The @xml:id attribute is used to document the unique reference.

element. This element is used to 'describe the languages, sublanguages, registers, dialects, etc. represented within a text' (TEI Consortium 2008, p. 46). The @ident (identifier) attribute supplies a language code defined by the IETF's (Internet Engineering Task Force) BCP 47 standard. [140]

Figure 5.20 provides a model for the header markup. In summary, there is information about the document/text (captured within the <fileDesc> section of the header) and there is information about the correspondence itself – the participants, their locations and the date of the letter (captured within the <profileDesc> section of the header). In the following chapter (chapter six) I will propose templates for modelling information about participants and locations, within the personography and placeography files.

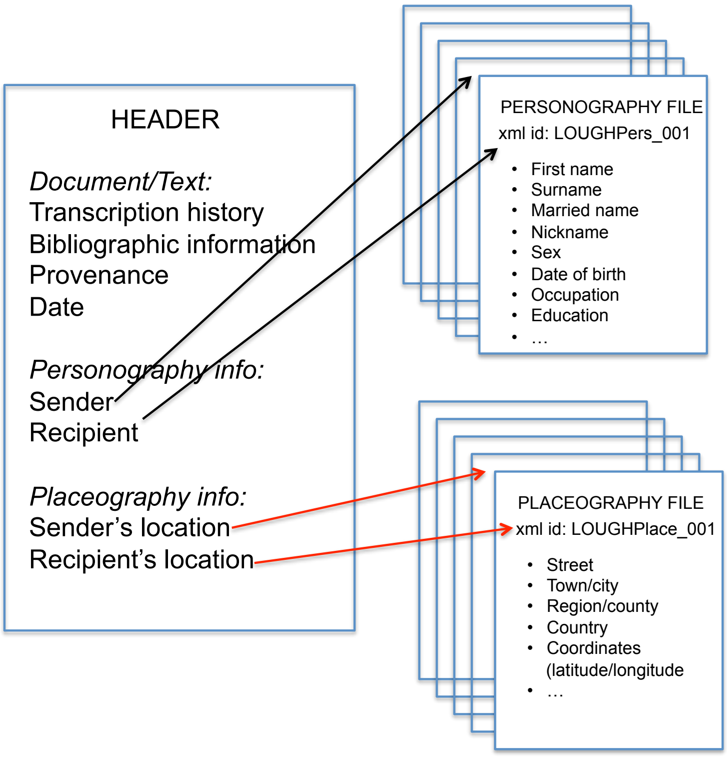


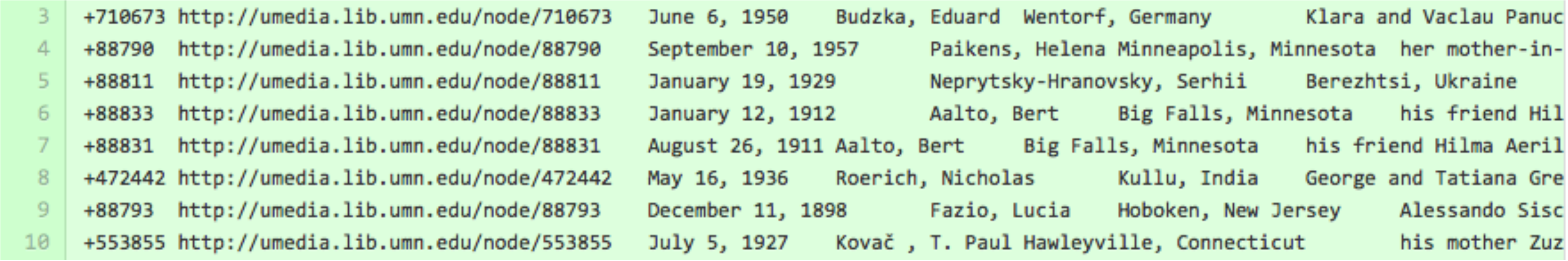
Figure 5.20: Modelling the header

[140] For more on language tags in XML see: *Tags for Identifying Languages* (2009) Available from: https://tools.ietf.org/html/bcp47 [Accessed 1 April 2015].

A small trial study: creating visualisations based on header markup

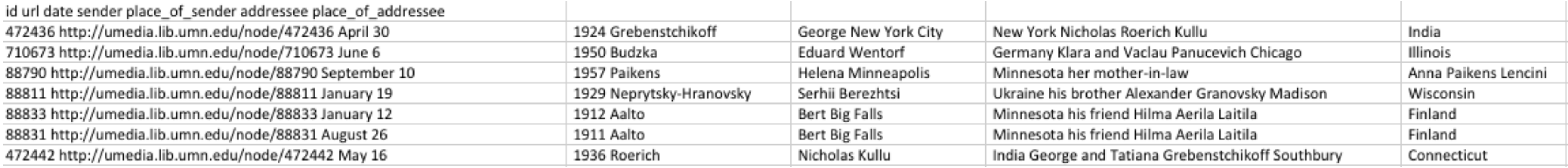
What follows is a brief summary of the work carried out as part of the DEM project to interconnect the metadata for <persName>, <placeName> and <date> relating to two emigrant letter collections that were mentioned in the literature review: the *Irish Emigration Database* (IED) – a collection of over 4,000 letters by Irish emigrants, held at the Mellon Centre for Migration Studies at the Ulster American Folk Park Museum in Omagh, Northern Ireland; and letters from the *Digitizing Immigrant Letters* (DIL) project at the Immigration History Research Centre at the University of Minnesota – a collection of 85 letters by migrants and their families in Europe and North America.

The first step was to standardise the metadata that is available for these two collections. An example of the raw metadata can be seen in Figure 5.21. One of the project partners, Luis Anke, Universitat Pompeu Fabra, converted this raw metadata into CSV format (see Figure 5.22). The standardised metadata was then passed to Peter Stadler, Universität Paderborn, to be made TEI compliant. This involved two stages: 1) converting the CSV file to TEI via OxGarage (http://www.tei-c.org/oxgarage/); 2) transforming generic TEI to <correspDesc> compatible TEI via XSLT.



```
+710673 http://umedia.lib.umn.edu/node/710673   June 6, 1950    Budzka, Eduard  Wentorf, Germany        Klara and Vaclau Panuc
+88790  http://umedia.lib.umn.edu/node/88790    September 10, 1957      Paikens, Helena Minneapolis, Minnesota  her mother-in-
+88811  http://umedia.lib.umn.edu/node/88811    January 19, 1929        Neprytsky-Hranovsky, Serhii     Berezhtsi, Ukraine
+88833  http://umedia.lib.umn.edu/node/88833    January 12, 1912        Aalto, Bert     Big Falls, Minnesota    his friend Hil
+88831  http://umedia.lib.umn.edu/node/88831    August 26, 1911 Aalto, Bert     Big Falls, Minnesota    his friend Hilma Aeril
+472442 http://umedia.lib.umn.edu/node/472442   May 16, 1936    Roerich, Nicholas       Kullu, India    George and Tatiana Gre
+88793  http://umedia.lib.umn.edu/node/88793    December 11, 1898       Fazio, Lucia    Hoboken, New Jersey     Alessando Sisc
+553855 http://umedia.lib.umn.edu/node/553855   July 5, 1927    Kovač , T. Paul Hawleyville, Connecticut        his mother Zuz
```

Figure 5.21: Example of the raw metadata from the DIL project, University of Minnesota



id url date sender place_of_sender addressee place_of_addressee				
472436 http://umedia.lib.umn.edu/node/472436 April 30	1924 Grebenstchikoff	George New York City	New York Nicholas Roerich Kullu	India
710673 http://umedia.lib.umn.edu/node/710673 June 6	1950 Budzka	Eduard Wentorf	Germany Klara and Vaclau Panucevich Chicago	Illinois
88790 http://umedia.lib.umn.edu/node/88790 September 10	1957 Paikens	Helena Minneapolis	Minnesota her mother-in-law	Anna Paikens Lencini
88811 http://umedia.lib.umn.edu/node/88811 January 19	1929 Neprytsky-Hranovsky	Serhii Berezhtsi	Ukraine his brother Alexander Granovsky Madison	Wisconsin
88833 http://umedia.lib.umn.edu/node/88833 January 12	1912 Aalto	Bert Big Falls	Minnesota his friend Hilma Aerila Laitila	Finland
88831 http://umedia.lib.umn.edu/node/88831 August 26	1911 Aalto	Bert Big Falls	Minnesota his friend Hilma Aerila Laitila	Finland
472442 http://umedia.lib.umn.edu/node/472442 May 16	1936 Roerich	Nicholas Kullu	India George and Tatiana Grebenstchikoff Southbury	Connecticut

Figure 5.22: Standardised metadata

Through standardising the metadata relating to these two collections, making it TEI compliant (in line with what is proposed in this chapter), it was then possible for Niall O'Leary (Freelance IT Specialist) to create a range of visualisations, exploring aspects of migration such as the movement of migrants over time and letter writing networks. These visualisations used a range of open-source libraries, but were tailored specifically for spatial, temporal and personal attributes. The visualisations can be found on Niall O'Leary's blog,[141] and highlight how, even with minimal metadata, it is possible to create meaningful visualisations which allow the user to notice patterns within the data. The map shown in Figure 5.23, for instance (developed using the open source map library, Leaflet), gives a general overview of the distribution of correspondence from both the IED and DIL collections (the blue dots represent destinations while the red dots represent the origins of the letters); whilst the visualisation in Figure 5.24 shows letter writing networks – by clicking on an individual's name, it is possible to view all persons the individual wrote to, or received letters from.[142]

[141] Niall O'Leary, Freelance IT Specialist: http://development.nialloleary.ie/correspondence/correspondence.php

[142] For more information about the creation of these visualisations see Moreton et. al. (2014).

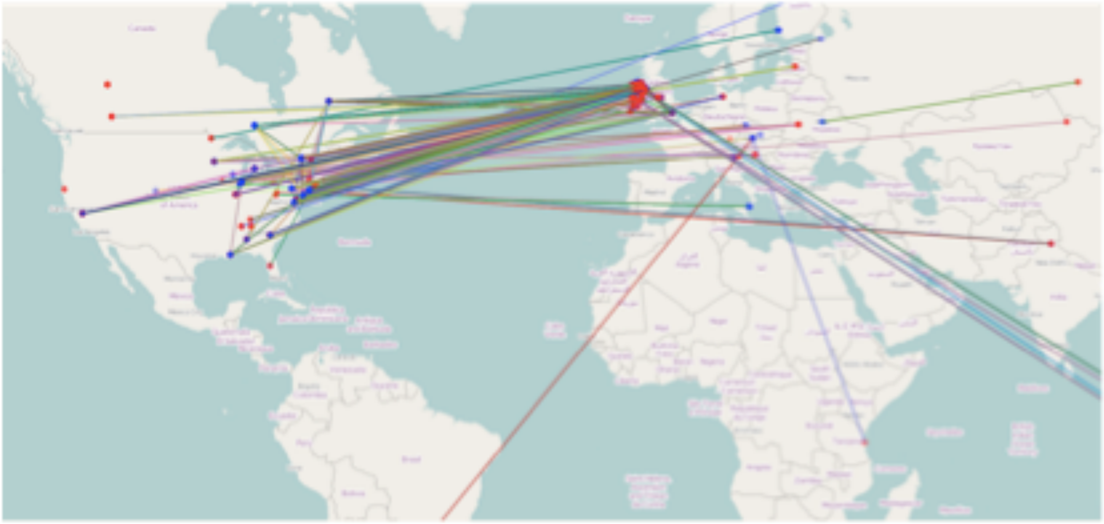


Figure 5.23: Visualisation 1 – distribution of correspondence

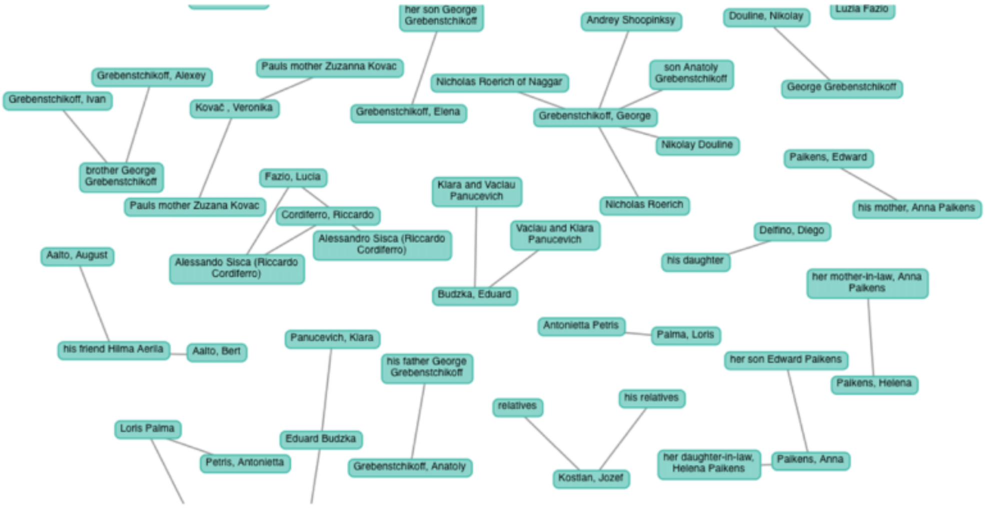


Figure 5.24: Visualisation 2 – letter writing networks

Future work

This chapter has focused on the description, organisation and categorisation of metadata relating to the emigrant letter, using TEI markup language. A lot more information could be included in the header – the level of detail will depend on factors such as time, budget and research requirements. Without a doubt, being able to categorise, classify and search emigrant letters based on sociobiographic information would be especially useful when working within and across large collections, and in the following chapter I will discuss how markup relating to participants and locations might be captured and organised in XML personography and placeography files, thereby enabling the user to carry out this type of search. Additionally, being able to categorise, classify and search letters based on their content would also be useful. However, agreeing on methods and tools for doing this, and agreeing on taxonomies which are meaningful across disciplines, is somewhat of a challenge and a lot more work needs to be done in this respect.

One particularly important feature of emigrant correspondence – and something which, arguably, distinguishes the emigrant letter from other types of letter – is their emotional content. As discussed in chapter one, text analysis software such as LIWC (*Linguistic Inquiry and Word Count*)[143] can provide a replicable way of identifying and capturing this sort of information, allowing users to narrow down their search to all letters with, for instance, a high frequency of positive or negative emotion words, while tools such as *WMatrix*[144] can also be used to identify statistically significant key semantic fields within each letter, potentially providing a useful overview of the letter content. Additionally, the *AntConc* and *Sketch Engine* findings from chapters two and three could be used to categorise letters based on their type/token ratio or the frequency of particular pronouns, for example, while the findings from chapter

[143] Pennebaker, J. W., Booth, R. J. and Francis, M. E. (2007) *Linguistic Inquiry and Word Count* (LIWC2007). Available from: http://www.liwc.net.

[144] Rayson, P. (2009) *Wmatrix*. Lancaster University. Available from: http://ucrel.lancs.ac.uk/wmatrix/.

four – if incorporated into the header markup – would allow users to search across letters for topics and themes.

Much of this information could be captured in the <profileDesc> element and/or the <sourceDesc> element. Within the <profileDesc> element there is the option to have a <textClass> (text classification) element which can be used to describe ‘the nature or topic of a text in terms of a standard classification scheme’ (TEI Consortium 2008, p. 45). Using the <keywords> element within <textClass> it is possible to categorise a text ‘by supplying a list of keywords which may describe its topic or subject matter, its form, date, etc.’ (TEI Consortium 2008, p. 47). Figure 5.25 provides an example of how the findings from chapter four might be captured in the <profileDesc> section of the TEI header, using Julia’s letter dated 25 January 1891 as an example, and Figure 5.26 provides an example of how some of the findings from chapters one to three (type/token ratios, pronoun usage, semantic fields etc.) might be captured within the <sourceDesc> section.

```
<profileDesc>
    <textClass>
        <keywords>
            <list>
                <item>salutation</item>
                <item>previous letters</item>
                <item>greeting</item>
                <item>homesickness and separation</item>
                <item>reunion</item>
                <item>health and illness</item>
                <item>recollections</item>
                <item>weather and seasons</item>
                <item>religion</item>
                <item>family and friends</item>
            </list>
        </keywords>
    </textClass>
</profileDesc>
```

Figure 5.25: Capturing topics in the <profileDesc> section of the TEI header (using *JLC*, 25 January 1891 as an example)

```
<sourceDesc>
    <p n="AntConc number of words">351</p>
    <p n="AntConc type token ratio">19.66</p>
    <p n="LIWC positive emotion">5.41</p>
    <p n="LIWC negative emotion">2.56</p>
    <p n="Wmatrix semantic fields">pronouns, existing, degree boosters, kin, getting and
    possession</p>
    <p n="LIWC pronoun I">7.12</p>
    <p n="LIWC pronoun you">4.27</p>
</sourceDesc>
```

Figure 5.26: Capturing textual information in the <sourceDesc> section of the TEI header (using *JLC*, 25 January 1891 as an example)[145]

To conclude, what this chapter has hopefully highlighted is the potential for working with header information – especially information embedded within <ct:correspDesc> relating to person (sender and recipient), location and date, without necessarily having access to the letter itself. One of the biggest challenges of working with historical emigrant correspondence relates to accessibility of letter collections and issues to do with intellectual property. It is often difficult to get access to collections (that is, the letters themselves – the body) and even more difficult to make collections freely available online – especially when working across disciplines and across cultures. However, by focusing on metadata about the letter (in other words, the header information) there are fewer barriers to overcome as regards interconnecting resources.

This is just the first step in terms of developing interoperable emigrant letter collections, but hopefully, with further work (and funding), more collections will be able to interconnect in the way described in this chapter.

[145] I would like to stress that Figures 5.25 and 5.26 are examples of the sort of information that might be captured in the TEI header allowing the user to narrow down their search based on text content and statistical information. I have not incorporated all of the findings from chapters one to four in the example markup shown here. Types of projection, for example, were not explored on a letter by letter basis, so it is not possible to say whether this particular letter (*JLC*, 25 January 1891) contains more projections of propositions, or proposals; however, this is perhaps an opportunity for further research. Additionally, exactly what information – what level of detail – and how that information should be organised within the header is a work in progress. I offer just one possibility in Figures 5.25 and 5.26.

CHAPTER SIX

Modelling personography and placeography information

Introduction

The previous chapter examined how metadata relating to emigrant letter collections can be organised and described within the header, using TEI markup language. The markup template that I proposed can be applied to different letter collections, thereby ensuring a certain level of consistency across projects. It is anticipated, however, that individual project teams will build on this basic template to create more refined headers that suit their specific requirements and research aims.

As discussed in chapter five, the metadata can be organised into three layers. The first layer – 'Document/Text' – captures information about the document and/or text itself, such as the transcription history and/or transcription practices, bibliographic information and details about the letter's provenance. The second layer – 'Personography' – captures information about whom the letter is from (the author and/or sender)[146] and whom the letter is to (the recipient). Finally, the third layer – 'Placeography' – captures information about the location of the sender and recipient. While chapter five suggested a method for organising and describing metadata relating to the first layer – 'Document/Text' – the present chapter examines how metadata relating to the second and third layers ('Personography' and 'Placeography') might be modelled in a formalised way, using TEI markup language.

[146] See chapter five regarding differences in meaning between the terms 'Author' and 'Sender'.

The formal description of personography and placeography information, relating to emigrant letters, needs to be consistent, systematic, and, ideally, agreed upon by the research community. With regard to this last point, in writing this chapter, I have drawn on workshop discussions from the 'Digitising Experiences of Migration' (DEM) project[147] to come up with TEI templates for personography and placeography metadata that are useful to scholars from a range of disciplines. Obviously, the more detailed the markup is the more control one has in terms of searching, sorting, filtering, suppressing and analysing the metadata later on; the level of granularity will depend on factors such as time, budget and research aims. As with chapter five, the present chapter proposes basic TEI templates that can be developed and refined by individual project teams at a later date. Although the templates capture a lot of metadata that is arguably relevant to any/all letter collections (the sender/recipient's name, for instance), I have tried to consider what it is about the emigrant letter that makes it different from other types of letter. In other words, I have tried to develop TEI markup templates that work specifically for emigrant letter collections, drawing out aspects, or themes, of migration that might be useful to scholars working with this type of data – information such as the emigrant's date of migration and the method of transportation, for instance.

In the following sections I will propose TEI markup templates for modelling personography and placeography information. As this chapter is proposing just one possible way of modelling this type of information, I will conclude by discussing the limitations of my proposal and where I think further work is needed.

[147] *Digitising experiences of migration: The development of interconnected letter collections*. Available from: www.lettersofmigration.blogspot.com.

Throughout the chapter the term ‘Participant/s’ is used to refer to any persons involved in the act of communication (the ‘Author’/‘Sender’ and ‘Recipient’), as well as any person/s mentioned within the letter content. I will refer to a range of letters from the Lough collection; however, to demonstrate the TEI template for ‘Personography’ I have used sociobiographic information relating to Elizabeth Lough and to demonstrate the TEI template for ‘Placeography’ I have used Elizabeth’s first letter, dated 7th March 1876.

Throughout my discussion the term ‘personography/placeography metadata’ refers to the tags that are assigned to information about people or places, thereby providing additional content and context to that information. For example, the <persName> tags in <persName>Elizabeth Lough</persName> would make explicit that the information contained between the tags is a person’s name, rather than, for instance, the name of a ship. The term ‘XML personography/placeography file’ is used to describe an XML file containing personography/placeography information that has been organised and structured in a TEI compliant format. Finally, the term ‘TEI markup template’ describes my proposal for modelling personography/placeography information using TEI markup language.

Modelling personography and placeography information

As outlined in chapter five, sometimes there is only limited information available as regards the participants involved in the act of communication. There may, for example, just be a first name for the sender and/or recipient – if that, and the address information may be limited to the name of a city, or a country. Sometimes, however, there is a lot of information (details to do with the sender/recipient’s family history, their occupation/s, their education, their places of residence and so on). Rather than capturing this information within the header itself, it is possible to use a pointing mechanism that directs the user from the header to a separate XML personography or placeography file, containing all information, and metadata, about that person or place.

To briefly recap from chapter five, within the <profileDesc> section of the header (see Figure 6.1) there is the <ct:correspDesc> element, which tells us whom the letter is from (the sender’s name and location), whom the letter is to (the recipient’s name and location), and when the letter was dated. The <ct:participant> element and the @role attribute tell us that what follows is information about either the "sender" or "recipient": <ct:participant role="sender"> and <ct:participant role="recipient">.

```
<profileDesc>
    <ct:correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-force-
    correspDesc">
        <ct:participant role="sender"
            <persName key="LOUGH_Pers0001">Elizabeth Lough</persName>
            <placeName key="LOUGH_Place0001">Winsted, Connecticut</placeName>
            <date when="1876-03-07"/>
        </ct:participant>
        <ct:participant role="recipient"
              <persName key="LOUGH_Pers0007">Elizabeth McDonald Lough</persName>
              <persName key="LOUGH_Pers0008">James Lough</persName>
              <persName key="LOUGH_Pers0002"/>Alice Lough</persName>
              <persName key="LOUGH_Pers0003"/>Anne Lough</persName>
              <persName key="LOUGH_Pers0004"/>Julia Lough</persName>
              <persName key="LOUGH_Pers0005"/>Mary Lough</persName>
              <placeName key="LOUGH_Place0006"/>Meelick, Queen’s County</placeName>
        </ct:participant>
    </ct:correspDesc>
</profileDesc>
```

Figure 6.1: Extract from the XML header file: the <ct:correspDesc> element

Focusing on metadata relating to the "sender", in Figure 6.1, the <persName> element tells us that the sender’s name is Elizabeth Lough while the

@key attribute provides a unique identifier (LOUGHPers_0001) that corresponds to an element within a separate XML personography file containing all details about Elizabeth Lough. Additionally, the <placeName> element tells us that Elizabeth's letter was sent from 'Winsted, Connecticut' while the @key attribute provides a unique identifier (LOUGHPlace_0001) that corresponds to an element within a separate XML placeography file containing all details about this location. Finally, the <date> element and @when attribute, embedded within <ct:participant role="sender">, captures when the letter was dated: 7th March 1876.

In what follows I will propose a way of modelling the information contained within the XML personography and placeography files (one file for each person and one file for each place),[148] using TEI markup language. Part one will focus on personography information, while part two will focus on placeography information.

Part one: personography information

Table 6.1 provides a summary of the sort of sociobiographic and sociohistorical information that the DEM project partners felt to be important when working with emigrant letters. As discussed in the previous chapter, it is very rare that all of this information will be available for any given letter. However, with a systematised and formalised way of capturing whatever information is available, it is easier to avoid duplications and improve interconnectivity and searchability across resources.

[148] Although it is possible to manage all personography and placeography information in two separate files (one file for all personography information and one for all placeography information), I have chosen to create XML files for each individual person and place as this makes it easier to organise, sort and manage the information at a later date.

Personography information *Sender/Recipient*	**Description**
First name	Sender/Recipient's first name/s. The name they were christened with rather than an abbreviated form.
Surname: married	Sender/Recipient's surname, including alternative spellings. For female authors it is possible to make a distinction between their 'married' surname and their 'maiden' surname, see below.[149]
Surname: maiden	Sender/Recipient's maiden name/s, including alternative spellings.
Nickname/s	All other names by which the Sender/Recipient was known.
Date of birth	Sender/Recipient's date of birth. If an exact date is not known, approximate dates can be provided, within a five or ten year span.
Date of death	Sender/Recipient's date of death. If an exact date is not know, approximate dates can be provided, within a five or ten year span.
Occupation/s	Occupation/s of the Sender/Recipient. Typically, this information will gleaned from the content of the letter.
Social status	Social status of the Sender/Recipient, primarily based on their occupation/s.
Education	Schooling of the Sender/Recipient: lower-rank mechanical schooling (MS); higher-rank grammatical schooling (GS).
Sex	Sex of Sender/Recipient: 1 = male; 2 = female; 9 = non-applicable; 0 = unknown.
Faith	Religious denomination of the Sender/Recipient.
Residences	The various residences of the Sender/Recipient, together with dates.
Relationships	Information about the relatives of the Sender/Recipient, specifically their spouse and parents – most other relationships can be derived from this information.
Migration	Information about when the Sender/Recipient emigrated, where they emigrated from/to, and the passage they took, together with method of transportation.
Date of marriage	Sender/Recipient's date of marriage, if known.

Table 6.1: Personography information to be modelled using TEI markup language

[149] These categories ('surname: married', 'surname: maiden') are somewhat biased towards European traditions. In some cultures, there is an important distinction between patronym and matronym which may be a relevant distinction to preserve.

Several of the categories listed in Table 6.1 are particularly difficult to determine and define, namely: ‘occupation/s’, ‘social status’, ‘education’ and ‘faith’. In terms of determining a participant’s occupation/s, there may be records that provide this kind of information (a marriage certificate, for example); more often than not, however, the only evidence available is found within the content of the letter itself, and even if an author does refer to an occupation they are likely to be describing what that occupation involves, rather than stating what that occupation is. Julia Lough, for instance, at no point refers to herself as a ‘seamstress’ or a ‘proprietor’, but instead writes about getting ‘all the sewing [she] can do’, or having ‘a shop of [her] own’, see *JCL, 18 January 1891* and *JLC, 4 June 1894*:

> *...I work*
> *home evenings and get*
> *all the sewing I can do*
> *but when I commence [sic passim] to*
> *get pay I will not take*
> *in sewing evenings as it is*
> *hard to work all the time*
> (*JLC, 18 January 1891*)

> *I am working and has*
> *got a shop of my own*
> *now on main street down*
> *stairs since June the*
> *place where I worked*
> *was smashing up so I*
> *started business myself*
> (*JLC, 4 June 1894*)

Similarly when trying to determine an author's faith, the researcher is often looking for clues in the letters: references to religious institutions, people, or events, for instance. Again, perhaps unsurprisingly, at no point does Julia write 'I am a Roman Catholic', which might sound strange in the context of writing to her family who are fully aware of her religious beliefs.[150] However, Julia does write about Ascension Day, members of her local church, and attending mass – see *JLC, 24 May 1893-94*, *JLC, 3 November 1889* and *JLC, 25 January 1891*:

> *This is a Holy day*
> *here Assension* [*sic passim*] *Day.*
> (*JLC, 24 May 1893-94*)

> *…Father Leo*
> *is our guide and director*
> (*JLC, 3 November 1889*)

> *I think it is*
> *dreadful to Stay from*
> *mass. The World is all*
> *very Well till our last day*
> *comes and then what*
> *have we but what little*
> *good we have done for our*
> *Souls.*
> (*JLC, 25 January 1891*)

[150] In other contexts and situations one can imagine Julia writing 'I am a Roman Catholic' – if writing to a new acquaintance, who has inquired about her religious beliefs, or if she has recently converted to Catholicism, for instance.

Even more problematic, and somewhat contentious, is determining the sub-categories for ‘social status’ and ‘education’. In terms of social status, Professor Miller categorises letters as LC (Lower Class), MC (Middle Class), or UC (Upper Class), based on criteria such as the participant’s occupation/s, their family history and their educational background; other factors to consider when determining social status might be whether the participant lived in rented accommodation or owned their own home. Establishing objective and replicable criteria, with clear boundaries, for determining social status is certainly problematic, not helped by the fact that, in some cases, the participant’s social status may have changed over time. (Julia Lough, for example, emigrates to America at just 13 years of age. Miller has categorised the Lough sisters as being from a lower class farming background, having had a minimal education. However, by the age of 24 Julia has her own shop employing several members of staff, arguably placing her in the middle class category.)

As with ‘social status’, the category ‘education’ poses similar issues. Sometimes there is sociohistorical evidence which reveals the type of schooling a particular participant had. More often than not, however, it is the participant’s occupation/s, their family history, and the language of the letter that provides clues as regards a participant’s schooling, as will be discussed later in this chapter.

In summary, determining and defining categories and labels for modelling sociobiographic and sociohistorical information is something that certainly requires ongoing work and discussion. This chapter proposes one possible method for organising information relating to people (and places), but it is by no means a definitive solution. Rather, my intention in this chapter is to initiate a cross-disciplinary discussion on how this type of information might be categorised, labeled and modelled in a way that is meaningful across the disciplines whilst at the same time allowing emigrant letter collections to potentially interconnect.

The information listed in Table 6.1 can be captured in a spreadsheet or database, such as that shown in Figure 6.2. As mentioned in the previous chapter, if information is stored in this way it is more manageable, with duplications or omissions in the data being easy to identify. A brief glance at the final row in Figure 6.2 (LOUGHPers_0030), for example, shows that there is very limited sociobiographic information relating to this entry. There is, in fact, just a surname. Gaps in the data – empty fields – can cause problems for programmers later on when it comes to manipulating the data and creating visualisations. Not only that, empty fields are ambiguous and do not tell the user whether the missing information is a mistake in the data, or whether the information in unavailable or unknown. Arguably, then, it is better to have a value for every field, even if that value is 'Unknown' or 'N/A' (non applicable). Furthermore, having a sense of what is missing, and why, is, arguably, valuable information in itself.

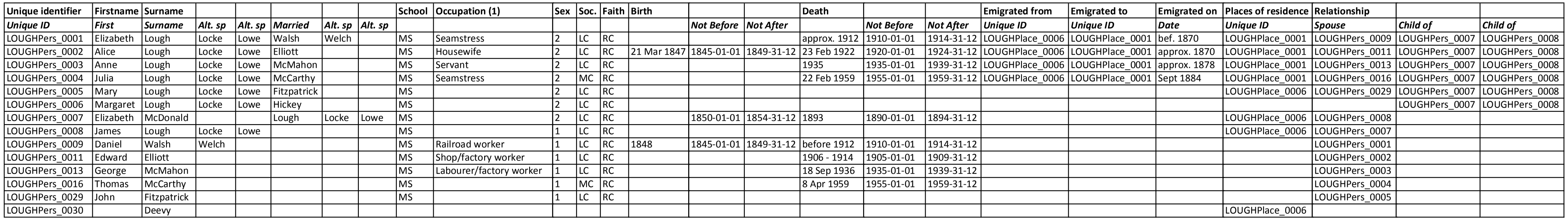
There are, however, issues that arise from using an 'Unknown' or a 'N/A' value – mainly technical rather than epistemological ones. Although using 'Unknown' or 'N/A' will work for string values[151] (such as a participant's name), they will not work for integers[152] where the database demands a numerical value for sorting purposes (for example, in the case of dates, or where numerical values

[151] The term 'string value' refers to groups of letters, including punctuation, digits, symbols and spaces (as opposed to numerical values) (Pine, 2003-2014).

[152] In most programming languages 'numbers without decimal points are called *integers*, and numbers with decimal points are usually called *floating-point numbers*, or more simply, *floats*' (Pine, 2003-2014).

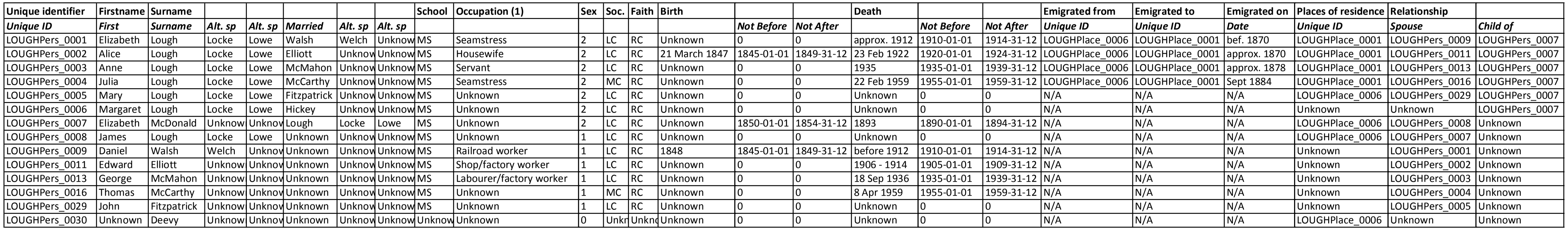
have been used to represent a participant's sex). Technically, a null value can be assigned to any field that is left blank (a null value '0' is a special value in database terms, which can be used in place of a number value). However, this does not satisfactorily get around the problem of partial dates, as will be discussed later in the chapter. In summary, when information is missing, it is preferable to assign a value to that field indicating why the information is missing, if doing so adds to our knowledge, if it is not at odds with the technical requirements of the system, and if it is not inaccurate. Ultimately, there is a difficult balancing act between giving the user as much accurate data as possible and actually presenting data that is not there. Figure 6.3, for example, provides the same information as Figure 6.2, but all empty fields have been given a value.[153] In instances where the field requires a string value, 'Unknown' is used when information (such as a participant's name, their schooling, or faith) is not known and 'N/A' is used when information is non applicable (the 'maiden name' column for male participants, for example). However, 'Unknown' and 'N/A' do not seem like entirely suitable values for empty fields in the 'Alternative Spelling' columns, and an 'Unknown' in one of the 'Relationship' columns seems somewhat ambiguous: does 'Unknown' mean that a participant's spouse is not known, or does it mean they did not marry in the first place? In many ways, agreeing on categories and labels for information that is missing, is just as important as agreeing on categories and labels for what is known. In the meantime, however, 'Unknown' and 'N/A' at least go some way to explaining empty fields.

[153] 'Unknown' means something different to 'N/A' and, therefore, carries a value in itself.



Unique identifier	Firstname	Surname						School	Occupation (1)	Sex	Soc.	Faith	Birth			Death			Emigrated from	Emigrated to	Emigrated on	Places of residence	Relationship		
Unique ID	*First*	*Surname*	*Alt. sp*	*Alt. sp*	*Married*	*Alt. sp*	*Alt. sp*							*Not Before*	*Not After*		*Not Before*	*Not After*	*Unique ID*	*Unique ID*	*Date*	*Unique ID*	*Spouse*	*Child of*	*Child of*
LOUGHPers_0001	Elizabeth	Lough	Locke	Lowe	Walsh	Welch		MS	Seamstress	2	LC	RC				approx. 1912	1910-01-01	1914-31-12	LOUGHPlace_0006	LOUGHPlace_0001	bef. 1870	LOUGHPlace_0001	LOUGHPers_0009	LOUGHPers_0007	LOUGHPers_0008
LOUGHPers_0002	Alice	Lough	Locke	Lowe	Elliott			MS	Housewife	2	LC	RC	21 Mar 1847	1845-01-01	1849-31-12	23 Feb 1922	1920-01-01	1924-31-12	LOUGHPlace_0006	LOUGHPlace_0001	approx. 1870	LOUGHPlace_0001	LOUGHPers_0011	LOUGHPers_0007	LOUGHPers_0008
LOUGHPers_0003	Anne	Lough	Locke	Lowe	McMahon			MS	Servant	2	LC	RC				1935	1935-01-01	1939-31-12	LOUGHPlace_0006	LOUGHPlace_0001	approx. 1878	LOUGHPlace_0001	LOUGHPers_0013	LOUGHPers_0007	LOUGHPers_0008
LOUGHPers_0004	Julia	Lough	Locke	Lowe	McCarthy			MS	Seamstress	2	MC	RC				22 Feb 1959	1955-01-01	1959-31-12	LOUGHPlace_0006	LOUGHPlace_0001	Sept 1884	LOUGHPlace_0001	LOUGHPers_0016	LOUGHPers_0007	LOUGHPers_0008
LOUGHPers_0005	Mary	Lough	Locke	Lowe	Fitzpatrick			MS		2	LC	RC										LOUGHPlace_0006	LOUGHPers_0029	LOUGHPers_0007	LOUGHPers_0008
LOUGHPers_0006	Margaret	Lough	Locke	Lowe	Hickey			MS		2	LC	RC												LOUGHPers_0007	LOUGHPers_0008
LOUGHPers_0007	Elizabeth	McDonald			Lough	Locke	Lowe	MS		2	LC	RC		1850-01-01	1854-31-12	1893	1890-01-01	1894-31-12				LOUGHPlace_0006	LOUGHPers_0008		
LOUGHPers_0008	James	Lough	Locke	Lowe				MS		1	LC	RC										LOUGHPlace_0006	LOUGHPers_0007		
LOUGHPers_0009	Daniel	Walsh	Welch					MS	Railroad worker	1	LC	RC	1848	1845-01-01	1849-31-12	before 1912	1910-01-01	1914-31-12					LOUGHPers_0001		
LOUGHPers_0011	Edward	Elliott						MS	Shop/factory worker	1	LC	RC				1906 - 1914	1905-01-01	1909-31-12					LOUGHPers_0002		
LOUGHPers_0013	George	McMahon						MS	Labourer/factory worker	1	LC	RC				18 Sep 1936	1935-01-01	1939-31-12					LOUGHPers_0003		
LOUGHPers_0016	Thomas	McCarthy						MS		1	MC	RC				8 Apr 1959	1955-01-01	1959-31-12					LOUGHPers_0004		
LOUGHPers_0029	John	Fitzpatrick						MS		1	LC	RC											LOUGHPers_0005		
LOUGHPers_0030		Deevy																				LOUGHPlace_0006			

Figure 6.2: Personography information captured in an Excel spreadsheet *



Unique identifier	Firstname	Surname						School	Occupation (1)	Sex	Soc.	Faith	Birth			Death			Emigrated from	Emigrated to	Emigrated on	Places of residence	Relationship	
Unique ID	*First*	*Surname*	*Alt. sp*	*Alt. sp*	*Married*	*Alt. sp*	*Alt. sp*							*Not Before*	*Not After*		*Not Before*	*Not After*	*Unique ID*	*Unique ID*	*Date*	*Unique ID*	*Spouse*	*Child of*
LOUGHPers_0001	Elizabeth	Lough	Locke	Lowe	Walsh	Welch	Unknow	MS	Seamstress	2	LC	RC	Unknown	0	0	approx. 1912	1910-01-01	1914-31-12	LOUGHPlace_0006	LOUGHPlace_0001	bef. 1870	LOUGHPlace_0001	LOUGHPers_0009	LOUGHPers_0007
LOUGHPers_0002	Alice	Lough	Locke	Lowe	Elliott	Unknow	Unknow	MS	Housewife	2	LC	RC	21 March 1847	1845-01-01	1849-31-12	23 Feb 1922	1920-01-01	1924-31-12	LOUGHPlace_0006	LOUGHPlace_0001	approx. 1870	LOUGHPlace_0001	LOUGHPers_0011	LOUGHPers_0007
LOUGHPers_0003	Anne	Lough	Locke	Lowe	McMahon	Unknow	Unknow	MS	Servant	2	LC	RC	Unknown	0	0	1935	1935-01-01	1939-31-12	LOUGHPlace_0006	LOUGHPlace_0001	approx. 1878	LOUGHPlace_0001	LOUGHPers_0013	LOUGHPers_0007
LOUGHPers_0004	Julia	Lough	Locke	Lowe	McCarthy	Unknow	Unknow	MS	Seamstress	2	MC	RC	Unknown	0	0	22 Feb 1959	1955-01-01	1959-31-12	LOUGHPlace_0006	LOUGHPlace_0001	Sept 1884	LOUGHPlace_0001	LOUGHPers_0016	LOUGHPers_0007
LOUGHPers_0005	Mary	Lough	Locke	Lowe	Fitzpatrick	Unknow	Unknow	MS	Unknown	2	LC	RC	Unknown	0	0	Unknown	0	0	N/A	N/A	N/A	LOUGHPlace_0006	LOUGHPers_0029	LOUGHPers_0007
LOUGHPers_0006	Margaret	Lough	Locke	Lowe	Hickey	Unknow	Unknow	MS	Unknown	2	LC	RC	Unknown	0	0	Unknown	0	0	N/A	N/A	N/A	Unknown	Unknown	LOUGHPers_0007
LOUGHPers_0007	Elizabeth	McDonald	Unknow	Unknov	Lough	Locke	Lowe	MS	Unknown	2	LC	RC	Unknown	1850-01-01	1854-31-12	1893	1890-01-01	1894-31-12	N/A	N/A	N/A	LOUGHPlace_0006	LOUGHPers_0008	Unknown
LOUGHPers_0008	James	Lough	Locke	Lowe	Unknown	Unknow	Unknow	MS	Unknown	1	LC	RC	Unknown	0	0	Unknown	0	0	N/A	N/A	N/A	LOUGHPlace_0006	LOUGHPers_0007	Unknown
LOUGHPers_0009	Daniel	Walsh	Welch	Unknov	Unknown	Unknow	Unknow	MS	Railroad worker	1	LC	RC	1848	1845-01-01	1849-31-12	before 1912	1910-01-01	1914-31-12	N/A	N/A	N/A	Unknown	LOUGHPers_0001	Unknown
LOUGHPers_0011	Edward	Elliott	Unknow	Unknov	Unknown	Unknow	Unknow	MS	Shop/factory worker	1	LC	RC	Unknown	0	0	1906 - 1914	1905-01-01	1909-31-12	N/A	N/A	N/A	Unknown	LOUGHPers_0002	Unknown
LOUGHPers_0013	George	McMahon	Unknow	Unknov	Unknown	Unknow	Unknow	MS	Labourer/factory worker	1	LC	RC	Unknown	0	0	18 Sep 1936	1935-01-01	1939-31-12	N/A	N/A	N/A	Unknown	LOUGHPers_0003	Unknown
LOUGHPers_0016	Thomas	McCarthy	Unknow	Unknov	Unknown	Unknow	Unknow	MS	Unknown	1	MC	RC	Unknown	0	0	8 Apr 1959	1955-01-01	1959-31-12	N/A	N/A	N/A	Unknown	LOUGHPers_0004	Unknown
LOUGHPers_0029	John	Fitzpatrick	Unknow	Unknov	Unknown	Unknow	Unknow	MS	Unknown	1	LC	RC	Unknown	0	0	Unknown	0	0	N/A	N/A	N/A	Unknown	LOUGHPers_0005	Unknown
LOUGHPers_0030	Unknown	Deevy	Unknow	Unknov	Unknown	Unknow	Unknow	Unknow	Unknown	0	Unkn	Unkno	Unknown	0	0	Unknown	0	0	N/A	N/A	N/A	LOUGHPlace_0006	Unknown	Unknown

Figure 6.3: Personography information captured in an Excel spreadsheet with ‘empty fields’ filled *

* Note that it was not possible to include all of the columns in these screenshots, but Figures 6.2 and 6.3 should give a sense of how personography information is captured in Excel spreadsheets

Another benefit of recording personography (and placeography) information within a spreadsheet/database is that it is relatively straight-forward, from a programming perspective, to then do a transformation which converts the content of the spreadsheet/database into TEI compatible XML format. As long as all of the required fields exist, a script can be written to write TEI directly from the spreadsheet/database. (While a spreadsheet/database is especially suited to structured data (such as the header information, discussed in chapter five), the advantage of XML is that it lends itself well to semi-structured data; however, provided a project knows the metadata fields it wants, and how to populate those fields, there is no problem in using a spreadsheet/database to capture personography and placeography information.)

I will now focus on how the information detailed in Figure 6.3 can be modelled in XML using TEI markup language.

```
<listPerson xmlns="http://www.tei-c.org/ns/1.0">
    <person xml:id="LOUGHPers_0001">
         <persName>
           <forename type="first">Elizabeth</forename>
           <surname type="married">Walsh</surname>
           <surname type="married" subtype="altSpelling">Welch</surname>
           <surname type="maiden">Lough</surname>
           <surname type="maiden" subtype="altSpelling">Locke</surname>
           <surname type="maiden" subtype="altSpelling">Lowe</surname>
           <addName type="nick">Liz</addName>
           <addName type="nick">Lizzie</addName>
       </persName>
       <birth notBefore="1845-01-01" notAfter="1849-31-12">Unknown</birth>
       <death notBefore="1910-01-01" notAfter="1914-31-12">approx. 1912</death>
       <occupation key="HISC#79510"
       ref="http://historyofwork.iisg.nl/detail_hiswi.php?know_id=20311&lang">seamstress</occupation>
       <occupation key="HISC#-1" ref="
       http://historyofwork.iisg.nl/detail_hiswi.php?know_id=19930&lang">householder</occupation>
       <socecStatus key="9" ref="http://historyofwork.iisg.nl/docs/hisco_hisclass12_book@_numerical.inc">lower-
       skilled</socecStatus>
       <education key="MS"/>
       <sex value="2"/>
       <faith key="RC"/>
       <residence notBefore="1870-01-01" notAfter="1874-31-12">Unknown between 1870 and 1875</residence>
       <residence key="LOUGHPlace_0001" notBefore="1875-01-01" notAfter="1879-31-12">1876-
       1878</residence>
       <residence notBefore="1880-01-01" notAfter="1914-31-12">Unknown between 1879 and 1912</residence>
       <listEvent>
           <event type="emigration" notBefore="1850" notAfter="1870" whereFrom="LOUGHPlace_0006"
           whereTo="LOUGHPlace_0001"><p>before 1870</p>
               <desc whereFrom="LOUGHPlace_0006" whereTo="LOUGHPlace_0004" <name
               type="transportation">Unknown</name></desc>
               <desc whereFrom="LOUGHPlace_0004" whereTo="LOUGHPlace_0014" <name
               type="transportation">Ship</name></desc>
               <desc whereFrom="LOUGHPlace_0014" whereTo="LOUGHPlace_0001" <name
               type="transportation">Unknown</name></desc>
           </event>
           <event type="marriage" notBefore="1870-01-01" notAfter="1874-31-12">before 1876</event>
       </listEvent>
    </person>
    <relationGrp>
       <relation name="spouse" mutual="LOUGHPers_0001 LOUGHPers_0002"/>
       <relation name="childOf" active="LOUGHPers_0001" passive="LOUGHPers_0007 LOUGHPers_0008"/>
    </relationGrp>
</listPerson>
```

Figure 6.4: TEI markup template for ‘Personography’, using sociobiographic information relating to Elizabeth Lough

Figure 6.4, above, provides a TEI markup template for personography information, using sociobiographic details relating to Elizabeth Lough for demonstration purposes. All metadata is contained within the element <listPerson>. Within <listPerson> there are two subgroups <person> (which contains metadata about the participant such as name, date of birth, date of death, occupation/s, social status, education, sex, faith, places of residence, date of emigration, date of marriage) and <relationGrp> (which contains metadata about the participant's relatives, namely their spouse and parents).

In the following sections, I will discuss <person> and <relationGrp> (plus all related elements) in turn. First, however, I will focus briefly on the relationship between the @key attribute in the header file (see the first <persName> element in Figure 6.1) and the @xml:id attribute in the personography file (see the <person> element in Figure 6.4, p. 227), as this is central to how the different XML files (the header, personography and placeography files) 'speak' to one another.

Within the TEI header (Figure 6.1), the <persName> element tells us that the sender's name is Elizabeth Lough while the @key attribute provides a unique identifier (LOUGHPers_0001) for Elizabeth. As discussed in chapter five, the @key attribute 'provides an externally-defined [project specific] means of identifying the entity (or entities) being named, using a coded value of some kind'.[154] In the case of the Lough collection, the @key value corresponds with three things:

1. A row within an Excel spreadsheet containing a person description. (The first cell in the first row of Figure 6.3, for example, contains the identifier for Elizabeth Lough – 'LOUGHPers_0001' – all subsequent information in that row (name, date of birth, sex, occupation/s, and so on) relates to that identifier).[155]

[154] *TEI Guidelines*. Available from: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.canonical.html [Accessed 20 July 2015].

[155] Although not discussed in this chapter, another benefit of assigning each participant a unique identifier is that it is then possible to mark any references to people/places within the letter content using the appropriate identifier, as follows: <persName key="LOUGHPers_0007">Mother</persName> or <placeName key="LOUGHPlace_0001">Winsted, Conn</placeName>. Although quite a labour intensive process, this has two benefits: 1) once marked-up it is possible to use visualisation tools to explore how, for instance, family networks stayed connected and informed about one another; and 2) it is possible to run a script which identifies all mention of, say, LOUGHPers_0001, to create a list of all the spelling variations and nicknames used to refer to this person – in this case,

2. The name of the XML file containing the person description for Elizabeth Lough (LOUGHPers_0001.xml). The XML file is created by extracting the relevant information from the Excel spreadsheet and converting this into a TEI compatible XML file.
3. The @xml:id value within the <person> element of the XML file described in (2) above (xml:id="LOUGHPers_0001").[156]

There are two things to note at this point. First, although project-specific @key values have been used with the Lough letters, these are not ideal for data interchange as '[n]o particular syntax is proposed for the values of the key attribute, since its form will depend entirely on practice within a given project'. This means that there is 'no way of ensuring that the values used by one project are distinct from those used by another'.[157] The TEI propose a more formal solution in the form of a @ref attribute with a tag URI scheme.[158] In instances where the personography data cannot be made freely accessible, because of copyright and intellectual property issues, it is possible to specify a relative URI. A relative URI points to a resource relative to its context (for example, a project-internal XML personography file, accessible within the edition).[159] The second

Elizabeth Lough. This list can then feed back into the personography file and inform later searches as well as the design of the search interface.

[156] Note that @xml:id is an XML attribute that has to occur on an element in an XML document; documents themselves cannot have an @xml:id.

[157] *TEI Guidelines*. Available from: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.canonical.html [Accessed 20 July 2015].

[158] 'A [URI] Uniform Resource Identifier is a compact string of characters that identifies an abstract or physical resource'. For information, 'A URI can be further classified as a locator, a name, or both. The term "Uniform Resource Locator" (URL) refers to the subset of URI that identify resources via a representation of their primary access mechanism (e.g., their network "location"), rather than identifying the resource by name or by some other attribute(s) of that resource. The term "Uniform Resource Name" (URN) refers to the subset of URI that are required to remain globally unique and persistent even when the resource ceases to exist or becomes unavailable' (Kindberg and Hawke 1998).

[159] Whereas '[a]n absolute identifier refers to a resource independent of the context in which the identifier is used…a relative identifier refers to a resource by describing the difference within a hierarchical namespace between the current context and an absolute identifier of the resource' (Kindberg and Hawke 1998). In other words, with a relative URI, part of the 'string' is missing

point to note is that in cases – such as the Lough collection – where personography files have been stored and catalogued in a project specific way, and do not have URIs assigned to them, a transformation can later be used to turn the individual XML files into HTML, with the unique identifier (for instance, LOUGHPers_0001) forming part of the URI. In other words, the unique identifier for Elizabeth Lough is a parameter that gets fed into the script that generates the HTML page.[160]

To summarise, there are different ways of pointing from the XML header file to the XML personography file, using the @key attribute or the @ref attribute. Figure 6.5 gives an example of the @key method and Figure 6.6 gives an example of the @ref method. The TEI guidelines suggest that while

> [t]he ref attribute should be used wherever it is possible to supply a direct link such as a URI to indicate the location of canonical information about the referent…[t]he key attribute is provided for cases where no such direct link is required: for example because resolution of the reference is carried out by some local convention, or because the encoder judges that no such resolution is necessary.[161]

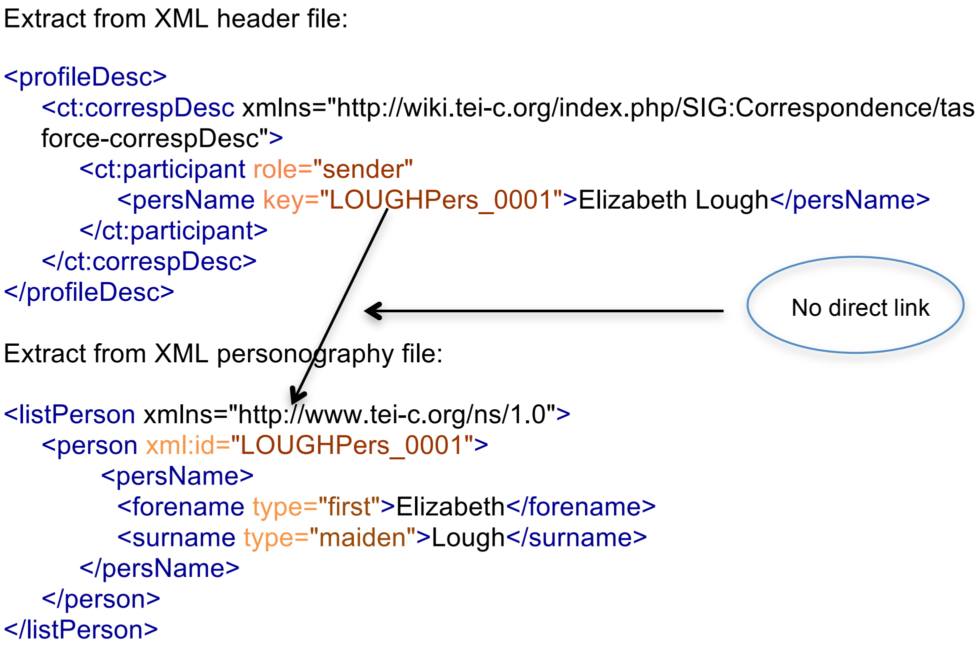
(i.e. a relative URI 'does not contain a fully qualified domain name and path, but instead contains just the path or a portion of the path' (https://yoast.com/relative-urls-issues/)). This missing information can be inferred, or is implied from the context.

[160] For demonstration purposes, the XML personography files for Elizabeth, Alice, Annie and Julia Lough have been assigned URIs (containing unique identifiers which correspond with those listed in column 1 of Figure 6.3) and made available on the web using the following username and password: development / Loughed-1900. The URIs are listed below. By clicking on the links (and entering the username and password) it is possible to access some of the personography information for each participant. I am indebted to Niall O'Leary (freelance programmer) for creating these links.

Name	**URI**
Elizabeth Lough	http://development.nialloleary.ie/lough/letters.php?xmlid=LOUGHPers_0001&letters_function=2
Alice Lough	http://development.nialloleary.ie/lough/letters.php?xmlid=LOUGHPers_0002&letters_function=2
Annie Lough	http://development.nialloleary.ie/lough/letters.php?xmlid=LOUGHPers_0003&letters_function=2
Julia Lough	http://development.nialloleary.ie/lough/letters.php?xmlid=LOUGHPers_0004&letters_function=2

[161] *TEI Guidelines*. Available from: http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ND.html [Accessed 20 July 2015].

Additionally, they recommend, where possible, using the @ref attribute with a tag URI scheme. Certainly, in terms of interconnecting resources, and avoiding duplication, this method is most desirable for emigrant letter projects, going forward.



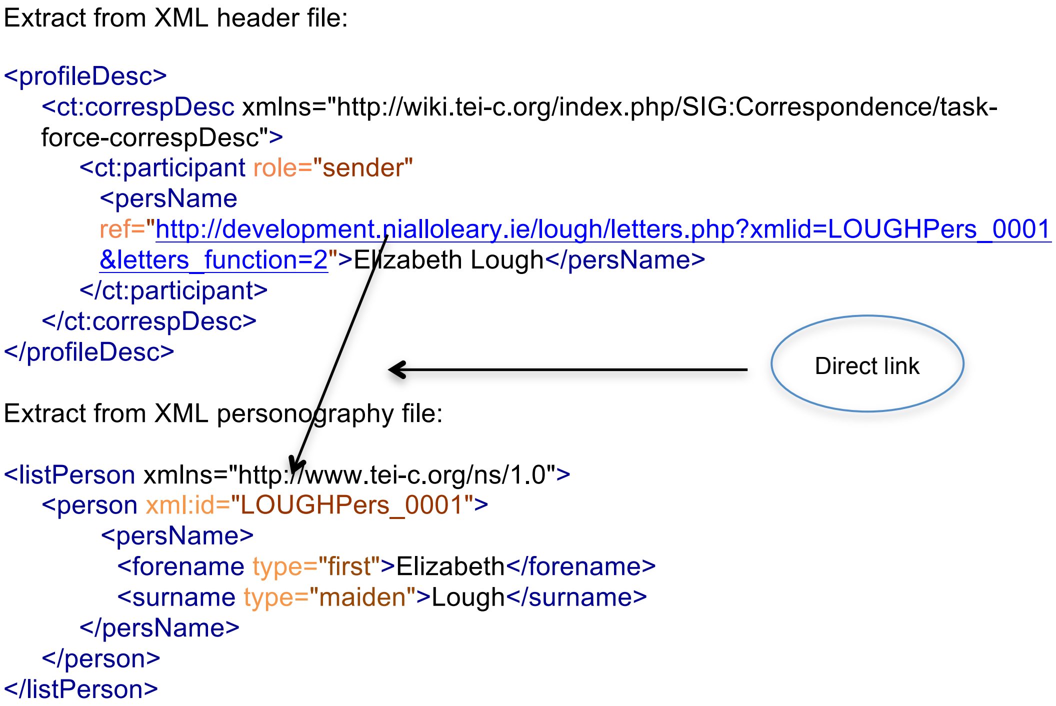
Extract from XML header file:

```
<profileDesc>
    <ct:correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-
    force-correspDesc">
        <ct:participant role="sender"
            <persName key="LOUGHPers_0001">Elizabeth Lough</persName>
        </ct:participant>
    </ct:correspDesc>
</profileDesc>
```

Extract from XML personography file:

```
<listPerson xmlns="http://www.tei-c.org/ns/1.0">
    <person xml:id="LOUGHPers_0001">
          <persName>
            <forename type="first">Elizabeth</forename>
            <surname type="maiden">Lough</surname>
        </persName>
    </person>
</listPerson>
```

Figure 6.5: Using the @key attribute to point to the XML personography file



Extract from XML header file:

```
<profileDesc>
    <ct:correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-
    force-correspDesc">
        <ct:participant role="sender"
          <persName
          ref="http://development.nialloleary.ie/lough/letters.php?xmlid=LOUGHPers_0001
          &letters_function=2">Elizabeth Lough</persName>
        </ct:participant>
    </ct:correspDesc>
</profileDesc>
```

Extract from XML personography file:

```
<listPerson xmlns="http://www.tei-c.org/ns/1.0">
    <person xml:id="LOUGHPers_0001">
          <persName>
            <forename type="first">Elizabeth</forename>
            <surname type="maiden">Lough</surname>
        </persName>
    </person>
</listPerson>
```

Figure 6.6: Using the @ref attribute to point to the XML personography file

The <persName> element

Figure 6.7 shows the metadata that is captured within the <persName> element of the XML personography file. Within <persName> are the elements <forename> and <surname>. The @type attribute is used to show whether the <forename> is "first" or "middle", and whether the <surname> is "married" or "maiden". Finally, the <addName> element captures any additional names (for example nicknames or pseudonyms) by which the participant is known – within the Lough letters, Elizabeth is often referred to, and refers to herself, as ‘Liz’ or ‘Lizzie’, for instance. Additionally, the @subtype attribute is used to capture any alternative spellings ("altSpelling") of the "married" and/or "maiden" names, as follows: <surname type="married" subtype="altSpelling">Welch</surname>. Note that @altSpelling is embedded within "married" or "maiden". This makes it explicit that a particular alternative spelling relates to either the married or maiden name.[162]

```
<persName>
   <forename type="first">Elizabeth</forename>
   <surname type="married">Walsh</surname>
   <surname type="married" subtype="altSpelling">Welch</surname>
   <surname type="maiden">Lough</surname>
   <surname type="maiden" subtype="altSpelling">Locke</surname>
   <surname type="maiden" subtype="altSpelling">Lowe</surname>
   <addName type="nick">Liz</addName>
   <addName type="nick">Lizzie</addName>
</persName>
```

Figure 6.7: The <persName> element

[162] An approach the TEI Correspondence SIG are taking is to wrap every name (married, maiden, pseudonym, etc) in its own <persName>. This frequently duplicates information (so I have chosen not to use this method for the Lough letters) – see, for example, the forename in:
<persName type="married"><forename type="first">Elizabeth</forename>
<surname>Walsh</surname></persName>
and
<persName type="maiden"><forename
type="first">Elizabeth</forename><surname>Lough</surname></persName>. However this method is more explicit when dealing with names in different languages.

Capturing all variations and alternative spellings of an author’s name is important when working with family history archives where the same name can sometimes be recorded in different ways. This is especially relevant when working with letters by minimally educated authors, who may spell their name phonetically (thereby creating a discrepancy between their sign off and the name given on their birth certificate or other official documents). Additionally, there may be mistakes or inconsistences within the official documents themselves that could cause researchers to miss relevant information when trawling through archives. When searching family history archives for information about the Lough family, for example, Miller found records filed under both ‘Locke’ and ‘Lowe’.

The <birth> and <death> elements

The <birth> and <death> elements capture details about the participant’s date of birth and date of death. If an exact date is known this can be recorded as follows: yyyy-mm-dd, as in <birth when="1860-01-01"/>. Partial dates, or unknown dates, are more problematic, especially for programmers, as they cannot be easily processed and sorted at a later date. A properly structured database should have complete dates (year/month/day) in date format; however, often a letter only contains part of the date, ‘7 March’, ‘March 1876’, or simply ‘1876’, for example. In some cases, the researcher may only have an approximate date – ‘between March and May 1876’, or ‘between 1870 and 1876’, for instance. As mentioned previously, if a numerical field is left blank it is automatically assigned the null value, '0', which, in this situation, is inappropriate. One possible solution is to use the @notBefore and @notAfter attributes within the <birth>

and <death> elements. For example, in the case of Elizabeth Lough, Miller's research suggests that Elizabeth died around 1912. This approximate date can be recorded (within a five year parameter)[163] using the @notBefore and @notAfter attributes, as follows: <death notBefore="1910-01-01" notAfter="1914-31-12">approx 1912</death>.

Elizabeth's date of birth is more difficult to determine. Miller's research suggests that Elizabeth was the first sister to emigrate to America. The earliest letter by Elizabeth, to her parents and sisters in Ireland, is dated 7 March 1876. It is not known whether other letters were sent home prior to this; however, the content of the 1876 letter reveals that Elizabeth arrived in America six years earlier:

> *...I no*
> *you all feel lonsom after her I did myself*
> *I could not help but cry to think she was*
> *left home an you Mother for I no the*
> *loss of a Father an Mother going on six*
> *years but you are still as fresh in my*
> *memory as the day I left home an always*
> *formost in my thoughts*
> (*ELC, 27 March 1876*)

Additionally, there is nothing to indicate how old Elizabeth was when she emigrated, although research by migration scholars suggests that the median age

[163] When determining what parameters to use (five years, ten years etc.) I had in mind the sort of searches users might want to carry out at a later date. Being able to narrow down a search as much as possible is, of course, preferable; however, the parameters need to work across letters and across collections. For the Lough letters, for the most part, it was possible to narrow down dates to within a five year span; however, a wider span may be required for other letter collections where specific, or partial, dates are not available.

for female migrants from Ireland to America in 1852 and 1921 was 21.2, as discussed in chapter one. Despite the lack of information, it is, nevertheless, possible – based on the letter content and research by migration scholars – to guess Elizabeth's approximate year of birth. (It is likely that Elizabeth was around 21 years old when she emigrated in around 1870, so her date of birth is likely to be around 1845-1850.) Using the @notBefore and @notAfter attributes makes it clear to users that the exact date of birth is not known. However, individual project teams would need to decide whether they want to make such approximations and it is often a case of getting a balance between creating searchable resources and keeping arbitrary manipulation of the data to a minimum. To summarise, within the spreadsheet shown previously (Figure 6.3), there is a column for 'Birth' that captures whatever information is available whether that is an exact date (27 March 1847) or a partial date (1848). This column contains string values, so it can accommodate values such as 'before X' or 'approx. Y'. The next two columns give a 'not before' and 'not after' numerical value, within a five year parameter. This means that when searching the data later on it will be possible for users to pull out all letters whose authors were born or died within a particular five-year span. There is, however, a problem with partial birth/death dates such as 'before 1912' as these do not give any indication as to how long before – five years, 20 years, or more – making it difficult to assign any parameters.

The <occupation> and <socecStatus> elements

When classifying data or texts, where possible, the TEI recommends taking keywords from a recognised source. The *Old Bailey Corpus*,[164] containing proceedings of the Old Bailey from 1674-1913, for example, categorises participants based on their occupation/s using the Historical International Standard Classification of Occupations (HISCO),[165] while a participant's social status is determined using HISCLASS (a social class scheme based on HISCO). A similar method can be applied to emigrant letter collections.[166]

The occupational titles (and codes) within the HISCO database come from various historical documents from around the world, dating from the 17th to 20th century. The full list of occupations can be accessed at: http://historyofwork.iisg.nl/list_hiswi.php?step=0&publish=Y. The HISCO database has 'a tree-like structure with 9 "major" groups, 76 "minor" groups, 296 "unit" groups and 1675 "micro" groups. The "leaves" of the tree are formed by the ten-thousands of occupational titles that fall under these 1675 groups' (http://hisco.antenna.nl/major.phtml).[167]

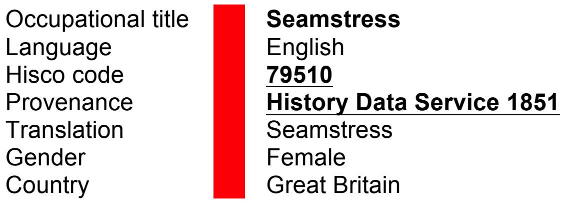
Taking Elizabeth Lough as an example, previous research by Miller shows that Elizabeth worked as a seamstress when she first emigrated to America. She then married and became a full-time mother. A search for the term 'seamstress', within HISCO, produces the following return:

[164] The *Old Bailey Corpus*. Available from: http://www.uni-giessen.de/oldbaileycorpus/ [Accessed 20 July 2015].

[165] History International Standard Classification of Occupations (HISCO). Available from: http://historyofwork.iisg.nl/ [Accessed 20 July 2015].

[166] HISCLASS. Available from: http://historyofwork.iisg.nl/list_pub.php?categories=hisclass [Accessed 20 July 2015].

[167] An overview of the HISCO structure can be found here: http://historyofwork.iisg.nl/list_pub.php?categories=hisco.



Occupational title	**Seamstress**
Language	English
Hisco code	**79510**
Provenance	**History Data Service 1851**
Translation	Seamstress
Gender	Female
Country	Great Britain

Figure 6.8: Search output for ‘seamstress’ in HISCO

The ‘Hisco code’, together with a URI link to the HISCO search output page, providing full reference information, can be incorporated into the markup, using the @key and @ref attributes, as follows:

```
<occupation key="HISC#79510"
ref="http://historyofwork.iisg.nl/detail_hiswi.php?know_id=20311&lang">seamstress</occupation>
<occupation key="HISC#-1" ref="
http://historyofwork.iisg.nl/detail_hiswi.php?know_id=19930&lang">householder</occupation>
```

Figure 6.9: The <occupation> element

Looking at Figure 6.9, the markup states that Elizabeth had two occupations. First she was a seamstress (the @key attribute points to the relevant ‘Hisco code’ for ‘seamstress’: HISC#79510, while the @ref attribute points to the URI, providing reference information relating to that ‘Hisco code’); and she was a householder. The term ‘householder’ is somewhat ambiguous; however, as the term ‘housewife’ did not produce any matches within HISCO the closest labels I could find were ‘householder’ or ‘mother’, both of which are viewed as a ‘status’ rather than an ‘occupation’ and are therefore given the code HISC#-1.[168]

Arguably the term ‘seamstress’ is not without its problems. The *OED* defines ‘seamstress’ as a ‘needlewoman whose occupation is plain sewing’. Indeed, Elizabeth may have been called and called herself a seamstress at the

[168] In HISCO, ‘[i]f a title containing status information gives no occupational information, it is given the appropriate STATUS code along with the HISCO code -1 or -2’ (http://historyofwork.iisg.nl/status.php?int02=11).

time, but one could question the appropriateness of using such a gender-biased term today. The term 'seamster' might be better – defined by the *OED* as 'a person whose occupation is sewing, esp. the mending and making of garments; a tailor'; however neither 'seamster' nor 'tailor' drew any search results in HISCO. The term 'dressmaker' does exist in HISCO however this occupation would move Elizabeth into a higher social grouping – that of 'skilled worker' – which may not accurately reflect Elizabeth's circumstances. It is possible to add new occupational titles to the HISCO database and there is certainly a case for evaluating the appropriacy of the HISCO categories/labels that are used to describe the various roles (emigrant) women lived and experienced; establishing categories which accurately describe those roles is perhaps an area for further research.

In terms of classifying participants based on their social status, the HISCLASS scheme, mentioned previously, offers one possible solution. In 2010, a file was created by van Leeuwen and Maas, Universiteit Utrecht,[169] which recodes HISCO (occupational information) into HISCLASS (class information), according to a procedure that is fully outlined in *HISCLASS: A historical international -social class scheme* (2011). In summary, there are 12 social status categories in total. Each occupation code listed in HISCO (except for two problematic cases, as explained in footnote 168) is assigned to one of these broad social class categories (1 = higher managers; 2 = higher professionals; 3 = lower managers; 4 = lower professionals and clerical sales; 5 = lower clerical and sales; 6 = foremen; 7 = skilled workers; 8 = farmers; 9 = lower skilled workers; 10 = lower skilled farm workers; 11 = unskilled workers; 12 = unskilled farm

[169] The file can be accessed at:
http://historyofwork.iisg.nl/docs/hisco_hisclass12_book@_numerical.inc.

workers). The occupation ‘seamstress’ (HISC#79510), for instance, is assigned to category 9 (lower-skilled). As ‘householder’ is defined as a status, rather than an occupation (i.e. it is one of the ‘problematic cases’), it is undefined in terms of social class. This social class information can be incorporated into the markup as follows:

```
<socecStatus key="9"
ref="http://historyofwork.iisg.nl/docs/hisco_hisclass12_book@_numerical.inc">lower-
skilled</socecStatus>
```

Figure 6.10: The <socecStatus> element

The @key attribute, in Figure 6.10, points to the relevant HISCLASS code for ‘lower-skilled’; while the @ref attribute points to the URI providing details of the full HISCLASS classification scheme.

The <education> element

For Fairman, membership of ‘rank’ or ‘class’ predicts a child’s schooling, of which, he argues, there are two comparatively distinct varieties: 1) lower-rank, mechanically-schooled (MS) writing; 2) higher-rank, grammatically-schooled (GS) writing (see Fairman 2008; 2012).

One possible way of determining a participant’s schooling, therefore, is to first identify their occupation (in accordance with the HISCO database), which will, in turn, correspond with one of the 12 HISCLASS categories for determining social class, which will, in turn, allow the user to predict (albeit tentatively) a participant’s probable schooling: ‘MS’ or ‘GS’. If, however, details about a person’s occupation/class are not known, then it is necessary to look for clues in the letter itself – the participant’s writing. Fairman – looking at 2,000 letters dating from the Late Modern period (since 1700) by artisans and the labouring poor – identifies several key differences between the writing of lower-rank mechanically-schooled writers and that of higher-rank grammatically-schooled writers, which can be summarised as follows:

1. While higher-rank GS writers tended to use longer latinate words, lower-rank MS writers tended to use longer stretches of monosyllabic units (Fairman points to contemporary quotes about the 'vulgar' lower-rank use of monosyllabic language). Additionally, lower-rank MS writers were more likely to spell words phonetically.
2. While higher-rank GS writers tended to shun formulas, lower-rank MS writers used more phrasal verbs and formulaic language (i.e. their semantic units stretched beyond single words).
3. Parataxis versus hypotaxis: lower-rank MS writers were less likely to use subordinate clauses (finite and non-finite) before the main verb than higher-rank GS writers. Any pre-main-verb clauses that they did write tended to be 'if' structures (as in 'if you do this, [then] that will happen'). Another feature of lower-rank MS writers was their use of chaining (as in 'I hope you are well, and John and Mary').
4. Higher-rank GS writers were more likely to use punctuation – and to use it consistently – than lower-rank MS writers.

There are, clearly, problems with determining a participant's schooling based solely on their writing. While children learn writing through some type of situation involving formal, purposeful teaching and/or learning (through a school, tutor, or parent, for instance), adults learn writing through autodidacticism, which, for lower-rank participants, can add any amount of features of, and confusions with, higher-rank grammatical schooling. Arguably, a third category – 'Unknown' – may be preferable in instances where a participant's occupation/social class (and, therefore, probable schooling) is not known.

Information about a participant's schooling can be captured in the markup as follows: <education key="MS"/>. The @key attribute value 'MS' tells us that Elizabeth Lough was minimally schooled as opposed to 'GS' (grammatically schooled) and points to a spreadsheet entry that provides a full definition of 'MS'. At present, within this spreadsheet for 'schooling', there are just three entries: 'MS', 'GS' and 'Unknown'; however, more refined categories could be added at a later date.

The <sex> element

The sex of a participant can be recorded using the @value attribute as follows: <sex value="2"/>. 'The <sex> element carries a value attribute to give the ISO 5218:1977 values (1 for male, 2 for female, 9 for non-applicable, and 0 for unknown)' (TEI Consortium 2008: 409).

The <faith> element

The faith of a participant can be recorded using the @value attribute as follows: <faith key="RC"/>. The @key attribute points to a spreadsheet entry which provides a full definition of 'RC' (Roman Catholic). In this spreadsheet I have based my categories for religious denomination on Miller's background research (RC = Roman Catholic; P = Presbyterian; PR = Protestant (Church of Ireland); E

= Episcopalian (Church of Ireland), and so on). Other categories – or sub-categories – can be added to this spreadsheet as more letter collections come on board.

The <residence> element

The movement of migrants within the New World is of special interest to migration scholars, allowing patterns of chain migration to be observed. In the case of the Lough sisters, for example, Elizabeth was the first sister to emigrate, preceded by her aunt and uncle – George and Anne Burke, who may have paid for Elizabeth's passage tickets. Younger siblings Alice, Annie and Julia followed later, most probably aided by Elizabeth. Capturing a participant's places of residence over their lifetime can be done using the <residence> element as follows: <residence key="LOUGHPlace_0099" notBefore="1875-01-01" notAfter="1879-31-12" />. Again, quite often exact dates are not known as to when a participant resided in a particular location so the @notBefore and @notAfter attributes can be used to give approximate timeframes. This is not ideal, however. For example, Elizabeth's first letter dated 1876 and all subsequent letters were sent from 'Winsted'. Elizabeth is referred to in letters by her siblings between 1878 and 1912 – with no mention being made of her moving house – however, the last letter we have by Elizabeth is dated 31 January 1877. In other words, although it is very likely that Elizabeth lived in Winsted from the time she emigrated in around 1870 to her death in around 1912, there is no concrete evidence regarding her places of residence between the periods 1870-1876 and 1878-1912. Again, this leads to the problem of what to do with missing information and whether to represent what is missing, or not. The general consensus amongst the TEI community is simply to delete elements or attributes for which the content is unknown as this facilitates processing; however, with emigrant letter collections in particular, I would argue that what is missing is just as important to model as what is there, as absences are an important part of the whole picture.

The <listEvent> element

A significant aspect of the emigrant's history is the event of migration itself (the date on which the participant migrated, where they migrated from/to, and the method of transportation used). The most obvious place to capture this information seems to be within the <event> element embedded within <listEvent>, as follows:

```
<event type="emigration" notBefore="1850" notAfter="1870" whereFrom="LOUGHPlace_0006"
whereTo="LOUGHPlace_0001"><p>before 1870</p>
   <desc whereFrom="LOUGHPlace_0006" whereTo="LOUGHPlace_0004" <name
   type="transportation">Unknown</name></desc>
   <desc whereFrom="LOUGHPlace_0004" whereTo="LOUGHPlace_0014" <name
   type="transportation">Ship</name></desc>
   <desc whereFrom="LOUGHPlace_0014" whereTo="LOUGHPlace_0001" <name
   type="transportation">Unknown</name></desc>
</event>
```

Figure 6.11: The <event> element

The @notBefore and @notAfter attributes are used to give an approximate date of migration (if the exact date is not known), while the @whereFrom and @whereTo attributes are used to show the start and end points of the migrant's journey.

There are different ways of capturing details of the journey itself. For example, short prose can be used within the <desc> (description) element to detail the passage took and the different methods of transportation used, as follows:

```
<desc>Elizabeth emigrated from Meelick, Queen’s County (now county Laois),
Ireland, to, most probably, Winsted, Connecticut. Although it is not known for
certain, it is likely (based on the passage her sister Julia took) that she first
travelled to Queenstown (County Cork) in order to get a ship to New York, before
travelling on to Winsted</desc>
```

Figure 6.12: Capturing information about the migrant’s journey using the <desc> element

I have chosen, however, to use tags to indicate the different points of the journey (from Meelick to Queenstown; from Queenstown to New York; and from New York to Connecticut) as well as the method of transportation at each stage, as this will aid searchability at a later date. It is easy to capture structured information within a spreadsheet/database, provided there are fields assigned for each tag (i.e. fields for ‘from’, ‘to’ and ‘transportation’). However, there is a problem with the <event> element in that it does not officially have the attributes @whereFrom and @whereTo, meaning that this part of my proposed template is not TEI compatible. An alternative is to use the <residence> element, which does allow the @whereFrom and @whereTo attributes, but then emigration is not listed as an ‘event’, which, I would argue, it should be. The distinction I am proposing here is that a participant’s movements within the New World (places of residence) are captured using the <residence> element, while the event of migration (date, locations, transportation) is captured using the <event> element. Although this means that the template is not entirely TEI compatible, a clear distinction can be made between places of residence and the event of migration.

Other significant events, such as marriage, can be listed within <listEvent> as follows: <event type="marriage" notBefore="1870-01-01" notAfter="1874-31-12">before 1876</event>. If the exact date is unknown, again, the @notBefore and @notAfter attributes can be used to give an approximate date.

Having looked at the elements embedded within <person>, I will now turn my attention to the final element: <relationGrp>.

The <relationGrp> element

As discussed throughout this thesis, a significant aspect of migration is to do with relationships and how individuals and families maintained those relationships over time and distance. By capturing information about a participant's relatives (specifically their spouse and parents – most other relationships can be deduced from this information) it is possible for a programmer to create visualisations showing family trees as well as letter writing networks. Details of a participant's spouse and parents can be captured using the <relation> element as follows:

```
<relationGrp>
    <relation name="spouse" mutual="LOUGHPers_0001 LOUGHPers_0002"/>
    <relation name="childOf" active="LOUGHPers_0001"  passive="LOUGHPers_0007
    LOUGHPers_0008"/>
</relationGrp>
```

Figure 6.13: The <relation> element

The @name attribute details the type of relationship (spouse, parent, etc.). It is also possible to model the nature of the relationship using the @mutual, @active and @passive attributes. While the @mutual attribute 'supplies a list of participants amongst all of whom the relationship holds equally', the @active attribute 'identifies the "active" participants in a non-mutual relationship' and the

@passive attribute ‘identifies “passive” participants in a non-mutual relationship’ (TEI Consortium 2008: 414). In the markup shown in Figure 6.13, we can see that LOUGHPers_0001 (Elizabeth) was married to LOUGHPers_0001 (Daniel Walsh) and she was the child of LOUGHPers_0007 (Elizabeth (MacDonald) Lough) and LOUGHPers_0008 (James Lough). The @mutual, @active and @passive attributes are potentially very useful when examining how language changes when participants are in a relationship that is equal/mutual, compared with a relationship that is ‘non-mutual’. Thinking back to chapter three, for example, there were noticeable differences in the use of projection structures depending on whether Julia was writing to her mother or her sibling. Having the capability to be able to search and analyse letters based on the type of relationship the letters contain would certainly be of use and interest to sociolinguists. Other types of relationship – friends, acquaintances, godparents, religious advisors – could be incorporated into the markup in a similar way; however, this would again depend on time and budget. In other words, it is possible to have a basic, required set of relationships and an optional, expandable set.

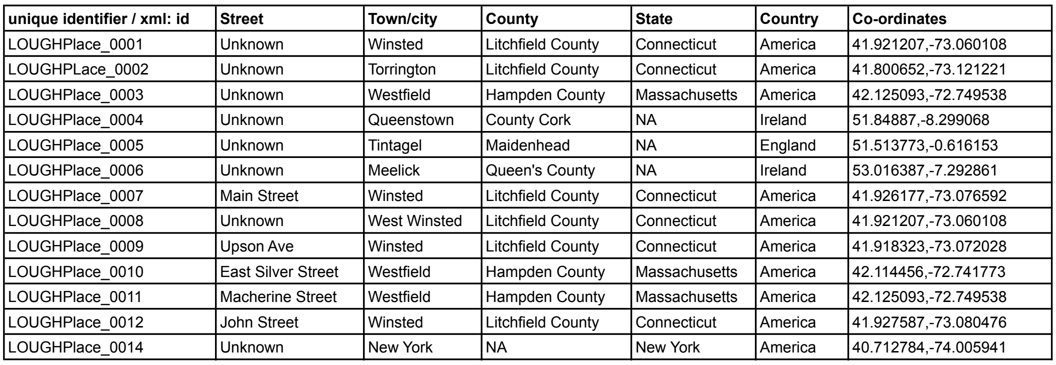
Part two: placeography information

Table 6.2 provides a summary of the key information that can be captured in relation to places (i.e. placeography information). This includes the house number, street name, village/town/city, region, country and GIS coordinates. Other information can be added to this list – the name of a building, whether it is a private or public space, for instance – however, the purpose of the TEI template for placeography information is to provide a basic skeleton which individual projects can build on and refine, depending on factors such as their data, budget and research aims.

Placeography information Sender/Recipient	**Description**
Street	The street name and number.
Village, town or city	The name of the village, town or city.
Region	The name of the region. This might be the State (for addresses in the US and Canada), or the County (for addresses in the UK and Ireland), for example.
Country	The name of the country.
GIS coordinates (latitude/longitude)	The GIS coordinates for the precise address (if known), or the village/town/city/region/country.

Table 6.2: Placeography information to be modelled using TEI markup language

As with information relating to person, information relating to place can be captured in a spreadsheet or database as shown in Figure 6.14.



unique identifier / xml: id	Street	Town/city	County	State	Country	Co-ordinates
LOUGHPlace_0001	Unknown	Winsted	Litchfield County	Connecticut	America	41.921207,-73.060108
LOUGHPLace_0002	Unknown	Torrington	Litchfield County	Connecticut	America	41.800652,-73.121221
LOUGHPlace_0003	Unknown	Westfield	Hampden County	Massachusetts	America	42.125093,-72.749538
LOUGHPlace_0004	Unknown	Queenstown	County Cork	NA	Ireland	51.84887,-8.299068
LOUGHPlace_0005	Unknown	Tintagel	Maidenhead	NA	England	51.513773,-0.616153
LOUGHPlace_0006	Unknown	Meelick	Queen's County	NA	Ireland	53.016387,-7.292861
LOUGHPlace_0007	Main Street	Winsted	Litchfield County	Connecticut	America	41.926177,-73.076592
LOUGHPlace_0008	Unknown	West Winsted	Litchfield County	Connecticut	America	41.921207,-73.060108
LOUGHPlace_0009	Upson Ave	Winsted	Litchfield County	Connecticut	America	41.918323,-73.072028
LOUGHPlace_0010	East Silver Street	Westfield	Hampden County	Massachusetts	America	42.114456,-72.741773
LOUGHPlace_0011	Macherine Street	Westfield	Hampden County	Massachusetts	America	42.125093,-72.749538
LOUGHPlace_0012	John Street	Winsted	Litchfield County	Connecticut	America	41.927587,-73.080476
LOUGHPlace_0014	Unknown	New York	NA	New York	America	40.712784,-74.005941

Figure 6.14: Placeography information captured in an Excel spreadsheet

For demonstration purposes, Figure 6.15 provides a TEI markup template for capturing placeography information, using Elizabeth Lough’s first letter, dated 7 March 1876; specifically, it models Elizabeth’s address information at the time of writing. Within the <place> element, the @xml:id value

("LOUGHPlace_007") corresponds with the @key value within the <placeName> element (under <ct:participant role="sender">) in the TEI header (shown in Figure 6.1).

```
<listPlace xmlns="http://www.tei-c.org/ns/1.0">
    <place xml:id="LOUGHPlace_007">
        <address>
            <street>Main Street</street>
        </address>
        <location>
            <settlement type="city">Winsted</settlement>
            <region type="state">Connecticut</region>
            <country>America</country>
            <geo>41.926177 -73.076592</geo>
        </location>
    </place>
</listPlace>
```

Figure 6.15: Placeography information organised using TEI markup

The <address> element

The <address> and <street> elements capture details of the house number (if known) and street name.

The <location> element

The <location> element captures the name of the village, town or city (using the <settlement> element and the @type attribute), the region or state (using the <region> element and the @type attribute), and the country (using the <country> element). Additionally, the <geo> element can be used to give the geographical coordinates for a specific address or region. Documenting the coordinates for a particular location is extremely useful when it comes to mapping the movement of migrants over time, or letter writing networks, for instance. Often the programmer will just want to pull out the WGS 84[170] coordinates, so presenting the coordinates in decimal format helps this process. It is possible to describe the notation and the datum used for geographic coordinates with the <geoDecl> element in the header, if a project feels it is necessary to do so (see: http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-geo.html).[171]

Discussion: limitations and future work

In this chapter I have looked at how information relating to people and places might be captured, labelled and organised in a structured and formalised way, using TEI markup language. If a similar system is applied across letter collections it becomes possible for resources to interconnect, allowing larger data sets to be explored and compared.[172]

As mentioned in the introduction, this chapter is by no means offering a definitive model for the markup of metadata relating to people and places, and there is certainly a lot more work to be carried out relating to the categorisation and classification of texts based on sociobiographic information. Rather, my aim

[170] The World Geodetic System. Available from: https://en.wikipedia.org/wiki/World_Geodetic_System [Accessed 20 July 2015].

[171] Additionally, with the help of a programmer, the information captured in the spreadsheet/database can be used to create a gazetteer, as follows:
1) A spreadsheet, such as that shown in Figure 6.14, is created which includes all information about the various places mentioned in a collection. The spreadsheet must include a unique identifier and a place name (expressed as, for example, 'City, State').
2) The letters themselves are then annotated for any references to places, using the <placeName> element, @key or @ref attribute, and the appropriate unique identifier.
3) An XSLT program converts the spreadsheet to a TEI gazetteer. This involves looking up the place names in geonames.org (this process can be automated) and extracting the relevant information from genoames.org into TEI structures.

[172] See, for example, the following projects which also use TEI to capture personography and placeography information:
Map of Early and Modern London. Available from: https://mapoflondon.uvic.ca and https://mapoflondon.uvic.ca/historical_personography.xml;
Colonial Despatches: The colonial despatches of Vancouver Island and British Columbia 1846-1871. Available from: http://bcgenesis.uvic.ca/places.xml;
and *UCLA Encyclopedia of Egyptology*. Available from: https://uee.ats.ucla.edu/welcome/.

has been to suggest basic templates which might be refined and built upon by individual project teams. Looking ahead, what is needed are more criteria that relate specifically to the topic of migration, and which draw out information about the living conditions and economic situation of the migrant, the institutions they were associated with, what private and public spaces they inhabited, and the social spheres the migrant belonged to – their context of community. However, even using the basic markup templates proposed in this chapter (and that of chapter five), when applied within and across letter collections, it is possible (with the help of a programmer) to explore a range of research question which might be difficult to explore otherwise – questions relating to, for example:

1) Patterns of letter writing

- The frequency of letter writing over time, including gaps in correspondence as well as any clusters of communication around particular times of the year: Christmas, New Year, or birthdays, for example.
- The intensity of exchange before, or after, significant life events, such as migration, marriage, the birth of a child, for instance.

2) Life stories

- A participant's life story – their occupation/s, marriage, children etc.
- The period of time from the date of migration to the date of marriage.
- Patterns in terms of the age, sex, faith, education etc. of the migrant at different periods and from different countries.

3) Patterns of chain migration

- Who migrated first, where did they settle, who followed?
- The passage the migrant took and the method of transportation used.
- The number of letters sent/received before migration takes place.

Although there are clear advantages to capturing and organising personography and placeography information in a formalised way (such as improved interconnectivity across resources as well as enhanced searchability), there are several drawbacks. First and foremost, it is time consuming and a relatively labour intensive process. In this chapter I propose capturing personography/placeography information in a spreadsheet or database, which can then be converted into TEI in-line with the templates being suggested. This would certainly save time (in my experience, it is easier to complete a spreadsheet than it is to complete TEI templates; and it is definitely easier to notice gaps in the data this way). Nevertheless, the process of populating the spreadsheet is still time consuming, mistakes can be made, and a programmer will be required to convert the spreadsheet information into TEI. An alternative method for capturing personography and placeography information is to set up a web-form. In many ways this is a better option than using, say, Microsoft Excel, as it constrains the choices the inputter has when completing the form. For information relating to 'sex', for instance, the inputter would just have four choices (1 = male; 2 = female; 9 = non-applicable; 0 = unknown), thereby reducing the possibility of error (although, of course, the wrong option could still be selected). Going forward, creating a central web-form, which anyone can contribute to, is perhaps the best way of capturing metadata across collections, cultures and countries, with people from different disciplines and different institutions being able to contribute metadata relating to their letter collections (both private and public) quickly and easily. This type of crowdsourcing approach would allow anyone with an interest in historical emigrant letter collections to contribute to the resource, helping to develop a fuller picture of exactly what letter collections are out there, where they are held, and the participants involved – the letter writes themselves, their lives and locations.

As gestured to several times throughout this chapter, further work is needed in certain areas. Specifically, I found it difficult to label the female emigrant in a way that truly reflected her occupation/s and status. As discussed previously, the labels in HISCO did not seem satisfactory somehow, and this is certainly an area which warrants further consideration and research. There were also epistemological and technical issues to do with representing vague or partial data, or, indeed, data that is missing. Again, although in this chapter I have given suggestions and workarounds for these issues, further work and cross-disciplinary discussion is required. The key questions are:

1. Should missing data be recorded and represented?
2. When is it okay to make an educated guess as to a person's age, date of migration, or faith, for instance?
3. How should missing information be recorded and represented (what labels should be used – 'Unknown', 'N/A', 'Uncertain' etc.)?
4. How can missing data be extrapolated and visualised in meaningful ways? (What will it tell us about the data, the topic of migration, or the migrant's experience?).

5. How can partial dates be represented in the markup? Are the @notBefore and @notAfter attributes a sufficient workaround?

Finally, there is the problem of how to represent ‘migration’ as an event in the author’s life. As discussed earlier, my belief is that ‘migration’ should be captured as a separate event, using the <event> element; however, according to the *TEI Guidelines*, it is not possible to use the @whereFrom and @whereTo attributes (crucial information relating to the process of migration) within the <event> element. It is arguably a balance between having metadata that is TEI compatible and having metadata that truly reflects, and brings out, information that is relevant and important to the subject of migration.

CONCLUSION

Once emigrant correspondence collections are digitised, computer programmes can order, search and make visible patterns within and across the letters. It becomes possible to step near and step back:

> Step Near…in that we can explore our research material in more detail and explore that detail in new, different ways…[t]he detail becomes available to a multitude of disciplines. And Step Back, in that we can observe our material from a distance, observe patterns that might not otherwise be easily visible, though they might have been postulated (Moreton et. al., 2014, p. 58).

In chapters one to four of this thesis I used corpus comparison tools (including *AntConc*, *Sketch Engine* and *Wmatrix*) to explore the language and content of the Lough letters. In chapters five and six I then proposed a method of TEI markup for capturing information about the letters – details about the document and/or text (the transcription history, bibliographic information and details about the letter's provenance) as well as sociobiographic information about the participants involved in the act of communication (that is, the author/sender and recipient), together with their various locations. My aim was to demonstrate how digital technologies can be used with (digitised) emigrant letter collections, offering new ways into the data, as well as allowing users to build on or challenge existing research.

Some scholars working in the digital humanities might argue that the corpus comparison tools described above make redundant the need for extensive markup. These tools can quickly analyse thousands of documents with the click of a button, while the process of markup is, to say the least, a labour intensive activity. I want to argue, however, that the text and content analysis tools used in this thesis do not by-pass the need for a system of markup for representing the sort of information categories described in chapters five and six. Tools such as *Wmatrix* can do certain kinds of textual analysis (but only some kinds), while markup enables a number of *con*textual and thematic analyses, allowing the user to extract all the letters from town X, or all the letters written in 1875, or all the letters which mention working conditions, for example. Indeed, without capturing such metadata, in a formalised and structured way, any corpus findings would have limited use and meaning. Patterns – in the language or content of the letters, and, for that matter, in the metadata itself – make more sense when viewed in their wider context – when viewed against the whole. The COBUILD corpus[173] is perhaps a good example of this issue. While a user can carry out different types of corpus inquiries, making potentially very interesting linguistic observations based on the search outputs, they are unable to say whether their observations have any correlation to socio-economic or -biographic variables. The lack of easily available metadata relating to the texts within the corpus means that a user has to do a considerable amount of hunting to find out who the speakers are – their age, sex, or nationality etc. For some scholars this is not a problem since the focus of their study is the text itself – independent of its context, while for others, myself included, language provides a window into the context of situation and

[173] The *Collins Cobuild Corpus* (2007-2013) Available from: http://www.collins.co.uk/page/The+Collins+Corpus [Accessed on 1 August 2015].

context of culture, potentially revealing how a person construes events and perceives the world, as well as revealing something about how, through language, identities are constructed, negotiated and performed.[174]

In chapters five and six of this thesis, I propose a possible method for describing, categorising and organising metadata relating to emigrant letter collections, using TEI markup language. The sort of information that I have captured helps users, across the disciplines, to narrow down their investigation, make comparisons within and across letter collections, and contextualise their findings. The TEI templates I propose for modelling 'Document/Text' (chapter five), and 'Personography' and 'Placeography' information (chapter six) are not necessarily finished products and I anticipate that different emigrant letter projects will use, critique and build-on these basic templates, documenting what did and did not meet their specific requirements, thus allowing the templates to be refined and developed over time. Projects may, for example, wish to capture information about the materiality of the letter, or whether the letter was written in pen or in pencil. (This connects emigrant correspondence research with the emerging discussion of the 'material letter' in other disciplines (see, for instance, Daybell and Hinds (2010) and Steen (1994).) They may wish to capture more detailed information about the authenticity and authorship of the letter (whether or not the letter was dictated, for example).[175] And they may also wish to record any watermarks, which might provide evidence of the existence of a papermaking business, and the trade and distribution of products, as well providing clues as to when a letter was written. Postmarks, too, can provide valuable information as to

[174] This is the thinking behind genre-based corpora, specialised corpora, local grammars, etc. The COBUILD project, by contrast, aimed to be a maximally mainstream and general corpus, valid for typical and ordinary English usage.

[175] On the subject of authenticity see Fairman (2000, p. 64) and Elspaβ (2012, p. 158).

where and when a letter was posted (especially useful when authors do not date their letters, nor name where they were written). Additionally, the postmark can provide insight into the time it took for a letter to reach its destination. It is not uncommon for an author to note the date on which they received a particular letter and if that letter has a postmark it is possible to calculate an approximate journey time. Finally, some projects may wish to capture information about enclosures or the amount/currency of any remittances, while others may wish to record any references to religious or secular institutions, which may be key to understanding how emigrant communities evolved and established themselves in the New World. The list could go on. While some scholars would argue that the features I have listed here are 'essential' to our understanding of the emigrant experience, others might describe this information as 'desirable'. When considering what information should be included in a system of markup for emigrant letters I tried to produce TEI templates that are broad enough to be applied across the disciplines. My criteria are ones that are likely to be robustly useful for a range of researchers, while pen versus pencil or similar distinctions are arguably more refined and delicate distinctions that can overlay the ones I propose in my templates without invalidating them. However, working within an interdisciplinary framework comes with its own unique set of problems and questions: why do we do interdisciplinary research and how do we understand good practice and research across the disciplines?

The process of deciding on information categories that are meaningful across the disciplines is one of the biggest challenges. In short, there needs to be a common language for talking about metadata relating to the emigrant letter. As already discussed in chapter five, the terms 'Sender' and 'Author', for example, are often used interchangeably by scholars working with emigrant letters, while the TEI Correspondence SIG makes a distinction between the two (the ‘Sender’ is the participant responsible for posting the letter; the ‘Author’ is the participant responsible for writing the letter). Similarly, with the terms ‘Recipient’ and ‘Addressee’, while the TEI Correspondence SIG use the term ‘Recipient’, other scholars prefer the term ‘Addressee’. Additionally, whereas the TEI Consortium use the terms ‘Personography file’ and ‘Placeography file’ to refer to master files containing metadata relating to people and places, archivists use the term ‘Authority file’ to describe the same thing.[176] Finally, as discussed in chapter six, there are issues to do with the sub-categorisation of very subjective, often contentious, terms such as ‘class’ and ‘education’: agreeing on the number of sub-categories and how to define and label those categories must be an interdisciplinary process. But interdisciplinarity takes varied forms, and is not unproblematic.[177] Nevertheless, I want to argue that ‘through examining the emigrant letter within a digital humanities framework it [is] possible for researchers from different disciplines to come together to discuss some of the problems and opportunities of working with [emigrant] correspondence collections, and to discuss best practice within and across disciplines’ (Moreton et. al., 2014, p. 59). Indeed, without interdisciplinary dialogue between myself, programmers, the TEI community, linguists and social-historians this thesis would not have been possible. But the dialogue needs to proceed in an agreed way, with clear research aims and outputs if the TEI templates set out in this thesis are to be honed and applied across emigrant letter projects.

[176] See Schuchard for a discussion regarding the ‘necessary complementarity’ between the archival and digital agendas (2002, pp. 62-63).

[177] As Griffin et. al. (2006) point out, there are few formal academic careers in interdisciplinary studies.

So far in my summing up I have suggested that while corpus comparison tools offer a possible way into emigrant letters, TEI markup helps to contextualise the resultant findings, allowing the user to compare and cross-check their observations against variables such as sex, occupation, location, class, date and so on. This suggests a divide between the two areas: text and content analysis tools are used to observe patterns in the language while TEI markup captures and organises metadata relating to the letter itself. To a certain extent this distinction is correct; but what I hope to have also demonstrated is that the findings from corpus comparison tools can be used to inform the creation of search criteria, which are later embedded into the TEI header.

For example, in chapter one, *Javascript Kit* was used to calculate the number of words and characters in each of the Lough letters. This information can then be captured within the <sourceDesc> element of the TEI header, allowing the user to explore, for example, whether the Lough sisters wrote lengthier letters at specific times of the year (Christmas or St Patrick's Day, for instance), or whether their letters tended to be longer or shorter towards the end of the correspondence cycle. Additionally, the *LIWC* results, when incorporated into the TEI header, offer at least one way of being able to identify those letters which focus on themes such as family, death, or work etc. In chapter two, *AntConc* was used to calculate the type/token ratios for each letter; again, this information can be captured within the <sourceDesc> section of the header, allowing users to search those letters that appear to be more formulaic in nature (with the same words being repeated), or those which appear to be more diverse (with a range of language being utilised).[178] Chapters two and three used *Antconc* and *Sketch Engine* to look more closely at, amongst other things, the use of pronouns and projection structures. The findings of these chapters – when incorporated into the TEI header – allow users to easily identify what seem to be more 'other' oriented letters (i.e. containing a significantly high frequency of the pronoun *you*), or more self-reflexive letters (i.e. containing a significantly high frequency of the pronoun *I*). Users might also wish to focus on letters that contain a high frequency of projections of propositions (exchanges of information, requiring a verbal response), or projections of proposals (exchanges of goods and services, requiring a non-verbal response – i.e. an action of some description).

While the corpus findings discussed so far are likely to be of particular interest to linguists, the ability to search letters for topics and themes is likely to appeal to a broader range of scholars from across the disciplines. Chapter four proposed a possible method for topic identification and analysis. It started with a close, personal reading of Julia's letters. From this reading, twenty-four main topics emerged. Each letter was then annotated for topics, <using angle brackets to show where a new topic begins and ends/>, and two of the twenty-four topics, 'Homesickness and Separation' and 'Recollections', were extracted and analysed using corpus comparison tools (in this case *Sketch Engine* and *Wmatrix*) to identify local grammars. Future research will involve establishing local grammars for all twenty-four topics and then testing to see whether those local grammars (specific words, phrases and patterns in the language) might indicate the thematisation of a particular topic in other letters. I certainly see potential for semi-automating the process of topic detection. Topics, once incorporated into

[178] Users wishing to explore formulaic language may wish to extract all letters with a type/token ratio of less than 10%, for instance.

the <profileDesc> element of the TEI header (as suggested in chapter five), will allow users to home in on a particular theme within and across collections, pulling out all references to ‘Family / Friends’, or ‘Migration’, for instance.

In sum, corpus comparison tools enable us to study thousands of letters at a time and to develop search criteria based on the findings from those analyses. In turn, this allows the end-user to narrow down their search in a greater variety of ways, pulling out all instances of letters which contain the topic ‘Health’ and/or which have a significantly high frequency of the personal pronoun ‘you’, for example. This, when coupled with other search variables relating to ‘Document/Text’, ‘Personography’ and ‘Placeography’ information, allows for some very sophisticated and creative search possibilities.

Despite the exciting opportunities offered by the digital humanities with regards to working with historical emigrant letter collections, there are several constraints which may hinder future research. The problem of interdisciplinarity has already been touched upon; however relating to this is the issue of differing attitudes, and laws, regarding copyright and intellectual property across disciplines, cultures and countries, which affects accessibility of resources and, I would argue, is the biggest barrier to interconnecting letter collections. As Honkapohja et. al. point out,

> although using open access transcriptions of original sources solves the problem of copyright for the texts, the copyright of manuscript images remains a problem. Since most manuscript repositories reserve the right to produce digital reproductions of their collections and charge significant fees for these reproductions, small projects in particular may be hard-

> pressed to obtain digital facsimiles even for their own use. Furthermore, since the repository that produced the reproductions owns the copyright for them, they cannot be freely published under an open access license (2009, p. 10).

Although Honkapohja et. al. are primarily referring to printed manuscripts, the same issues exist for handwritten ego-documents such as the personal letter. The only way around the problem, Honkapohja et. al. argue, is to 'work with repositories and persuade them to either digitise the manuscript material and to publish them under an open access license, or to allow scholars to photograph manuscript material themselves' (ibid.).

Sustainability of digital resources is also an issue, which Millett argues is exacerbated, in part, by 'the tendency of large electronic projects…to use complex custom-built software, which makes them particularly difficult to update' (2013, p. 46). Another problem, touched on earlier, relates to time. The markup process, as already mentioned, is labour intensive and time consuming. In chapters five and six I suggested that metadata relating to emigrant letter collections can be stored in a spreadsheet or a database making it easier to notice gaps and inconsistences in the data. A transformation can then be used to convert the spreadsheet or database information into TEI. This would certainly speed up the encoding process and make it more manageable, but it does not get around the fact that a human would still need to input and check the metadata in the first place. Another problem relates to cost. Of course, with a large budget the previous problem of time would be less of an issue: research assistants could be hired to input information into the spreadsheet, and programmers could be employed to do the transformations. However, funding is never limitless and projects need to make important decisions about where they want to put their money. A sophisticated markup with extensive granularity could be created allowing both specialist and non-specialist users to input metadata relating to a particular letter collection. Or, alternatively, all of the metadata could be gathered in one document and a skilled programmer could be hired to effectively sort out and manipulate that metadata. In other words, a project might invest money in having fully extensive markup, or they might choose to invest in intelligent programming.

What I would like to propose in this thesis is that time and money would be better invested in developing sophisticated webforms, which allow information categories (such as those described in chapters five and six) to be captured in an organised, formalised and structured way. Additionally, the use of webforms to capture metadata offers great potential in terms of crowd-souring and research-sourcing possibilities, allowing both specialists and non-specialists to contribute to a central resource of metadata relating to emigrant letters. The use of webforms potentially restricts the possibility for error. The user would have drop down lists for the various (TEI) elements and attributes, from which they would simply choose the appropriate option. This means that anyone (the general public with an interest in family history, archivists, librarians, heritage centres, scholars from across the disciplines) who discovers a letter collection can document that collection – and all related metadata – at a central site, by inputting the relevant 'Document/Text', 'Personography' and 'Placeography' information into a user-friendly webform. This central site would thus grow organically with a range of users, from a range of backgrounds, contributing to its development.[179] And there is potential here too for using crowd-sourcing methods to engage students and the public in the interpretative markup of topics, although this would be an area for future research.

It is worth mentioning at this point in the discussion the importance, too, of having reliable digital transcriptions. Issues relating to transcription practices were touched on in chapter five, however it should be stressed that without reliable transcriptions we do not have reliable data. As Fairman (*forth.*) points out, transcribing, or copying, is not easy, especially when working with unfamiliar language. Alter (2009: xxxv), discussing the problems of translating Psalms in the Old Testament, writes, '[i]t is a nettlesome truth about scribal transmission that any text copied by scribes from century to century accumulates errors over time. A copyist's eye can easily skip over a letter, a word or even a whole phrase'. He goes through several common mistakes with copying letters. In fact, Fairman argues that the task of converting written discourse from handwritten to printed mode, thus making it available for digitisation, is often best seen as 'transliteration' more than 'transcription' in as much as it is often necessary to learn an author's way of writing (*forth*.). (The author may, for example, intend to write one graph ('p') but produces something which resembles the form of another graph ('f'), for instance.) There is certainly a case for developing best practice guidelines for transcribing historical emigrant letters.

To conclude, I would like to outline future possibilities and opportunities for the digitisation, markup and analysis of emigrant letter collections. As mentioned previously, copyright and intellectual property issues pose the biggest

[179] A monitor/editor would be required to oversee such as project/resource.

barrier to interconnecting and accessing resources. However, working with metadata relating to letter collections (rather than the letters themselves) presents fewer issues. The metadata, I believe, offer a possible starting point for interdisciplinary research. Specifically, what is needed is agreed upon information categories that truly reflect and draw out important aspects of migration and the emigrant experience. This thesis puts forward one possible way of doing this. Through continued dialogue across the disciplines, the TEI templates that I have proposed for 'Document/Text', 'Personogarphy' and 'Placeography' can be refined and webforms created which allow users to record metadata relating to their collections in an easy and efficient way via a central website. This central hub would provide a much needed overview of existing collections (what resources exist and where), as well as providing excellent contextual information for anyone interested in migration history, allowing the user to explore questions such as:

- How many letters/collections are there?
- Where are those letters/collections housed?
- Who are the authors, senders and recipients (sex, age, occupation, class etc.)?
- Where are the participants located?
- Who (or where) is underrepresented or not represented at all?

Modelling information about the emigrants themselves as well as mapping their movements over time provides invaluable information about how emigrants and emigrant communities communicated, evolved, functioned and stayed connected. And much of this can be achieved through analysing the metadata. Furthermore, this metadata can be used to create meaningful visualisations which allow the user to notice patterns within the data (i.e. what is there) and, just as important, gaps (i.e. what is not there). Additionally the metadata relating to lots of letters can be stored and retrieved at once. The key question, then, is how to read this metadata. Obviously it is possible to read the metadata relating to each letter one at a time. And it might be possible to identify trends and patterns this way. Ultimately, though, very large datasets may prove hard to grasp in their entirety. Even a relatively small dataset – such as the Lough collection – can be difficult to analyse in this way. This is where data visualisation comes in. By processing the metadata into a graphical form it is possible to navigate the content more easily, notice trends and patterns, home in on specific stories and bring together content from many different sources. Although data visualisations do not necessarily provide answers to research questions, they give a good basis for further research, allowing the user to gain new insights into historical data, hopefully giving it a new resonance for today.

Finally, on the subject of data visualisation, I would like to argue that geospatial experts should be included in any discussions on how to capture and visualise information relating to locations (whether that is the location of the author, sender or recipient, or locations mentioned within the letter content). The movement of people over time is absolutely central to the emigrant story and finding new ways to capture and represent this sort of information must be at the heart of any digital emigrant letters project.

I would like to end with a final word about gaps in the data, and the ongoing question of how to represent what is not there – and, indeed, whether it is necessary to represent what is missing. In the introduction I discussed the idea of there being gaps in social history – the unheard voice of the subaltern, and gaps in language – 'standard' versus 'non-standard'. And arguably this notion of absence is a recurring theme in our understanding of the emigrant letter.[180] Time and time again, when reading the Lough letters, I was struck more by what is not said, than what is said: the stretches of silence between letters sent and received, the family members who are infrequently mentioned in the content, and the avoidance or omission of certain topics, for example. We need to find ways of identifying, understanding and representing what is missing as, arguably, silences and omissions are just as important as the thoughts and feelings that are committed to paper. As discussed in chapter six, there needs to be an agreed upon system for documenting absences (the labels that are used to indicate whether information is missing or unknown, for example). However, I would argue that capturing metadata – in the way that I propose in this thesis – is a crucial stage along the way to understanding what is there and what is absent.

[180] The issue of silences in emigrant correspondence is touched on by Richards (2010; 2006). See also Poland and Pederson (1998).

REFERENCES

Ädel, A. (2012) 'What I want you to remember is…' Audience orientation in monologic academic discourse. *English Text Construction. Special issue: Intersections of Intersubjectivity* 5 (5): 101–127.

Ädel, A. (2006) *Metadiscourse in L1 and L2 English*. Amsterdam: John Benjamins Publishing Company.

Akenson, D. H. (1993) *The Irish Diaspora: a Primer*. Toronto: P. D. Meaney Co.

Honkapohja, A., Kaislaniemi, S. and Marttila, V. (2009) Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora. In A. H. Jucker, D. Schreier and M. Hundt (eds.), *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29).* Amsterdam/New York: Rodopi.

Alter, R. (2009) *The Book of Psalms: A Translation with Commentary*. London: W. W. Norton & Company.

Amador-Moreno, C. P. and McCafferty, K. (2012) Linguistic identity and the study of Irish letters: Irish English in the making. *Lengua y Migración* 4 (2): 25-42.

Amador-Moreno, C., Corrigan, K.P., McCafferty, K., Moreton, E. and Waters, C. (2015, *forth*) Irish Migration Databases as Impact Tools in the Education and Heritage Sectors. In K. Corrigan and A. Mearnes (eds.), *Creating and Digitizing Language Corpora – Volume 3: Databases for Public Engagement*. London: Palgrave.

Atkinson, A. (1997) *The Europeans in Australia: A History, Volume One: The Beginning*. Oxford: Oxford University Press.

Atkinson, A. (2014) *The Europeans in Australia: Volume Three: Nation*. Sydney: University of New South Wales Press.

Atkinson, A. (2005) *The Europeans in Australia: A History, Volume Two: Democracy*. Oxford: Oxford University Press.

Auer, A. and Fairman, T. (2012) Letters of Artisans and the Labouring Poor (England, c. 1750-1835). In P. Bennett, M. Durrell, S. Scheible and R. J. Whitt (eds.), *New Methods in Historical Corpus Linguistics*. Narr: Tübingen. pp. 77-91.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Barton, H. A. (1990) *Letters from the Promised Land: Swedes in America, 1840-1914*. Minneapolis: University of Minnesota Press.

Bakhtin, M. M. (1986) *Speech Genres and Other Late Essays*. [Translated by Vern W. McGee.] Austin, Texas: University of Texas Press.

Bergvall, V. L., Bing, J. M. and Freed, A. F. (1996) *Rethinking language and gender research*. London: Longman.

Biber, D., Finegan E. and Atkinson, D. (1994a) ARCHER and its challenges: Compiling and exploring A Representative Corpus of Historical English Registers. In U. Fries, P. Schneider and G. Tottie (eds.), *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora, Zurich 1993*. Amsterdam: Rodopi.

Biber, D., Finegan, E., Atkinson, D., Beck, A., Burges, D. and Burges, J. (1994b) The design and analysis of the ARCHER corpus: A progress report [A Representative Corpus of Historical English Registers]. In M. Kytö, M. Rissanen and S. Wright (eds.), *Corpora across the centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25-27 March 1993* (Language and Computers. Studies in Practical Linguistics 11). Amsterdam and Atlanta: Rodopi.

Biber, D. (1995) *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Blegen, T. C. (1955) *Land of Their Choice: The Immigrants Write Home*. Minneapolis: University of Minnesota Press.

Brettell, C. B. and Hollifield, J. F. (2000) *Migration Theory: Talking Across Disciplines*. New York: Routledge.

Brinks, H. J. (1991) Impressions of the ‘Old’ World, 1848-1940. *European Contributions to American Studies* 20: 34-47.

Brinton, L. J. (2008) *The Comment Clause in English. Syntactic Origins and Pragmatic Development*. Cambridge: Cambridge University Press.

Bruce, S. U. (2006) *The Harp and the Eagle: Irish-American Volunteers and the Union Army, 1861-1865*. New York: New York University Press

Burnett, J., Vincent, D. and Mayall, D (1984) *The Autobiography of the Working Class. An Annotated, Critical Bibliography. Vol. 1: 1790-1900*. Brighton: Harvester Press.

Cancian, S. (2010) *Families, Lovers and their Letters: Italian Postwar Migration to Canada*. Manitoba: University of Manitoba Press.

Cannadine, D. (1998) *Class in Britain*. London: Penguin.

Connolly, C. (2001) Theorizing Ireland. *Irish Studies Review* 9 (3): 301-15.

Cano Aguilar, R. (1996) Lenguaje 'espontáneo' y retórica epistolary en cartas de emigrantes españoles a Indias. In T. Kotschi, W. Oesterreicher and K Zimmermann (eds.), *El Español hablado y la cultura oral en España e Hispanoamérica*. Frankfurt/Madrid: Vervuert/Iberoamericana. pp. 375-404.

Chafe, W. (1986) Evidentiality in English conversation and academic writing. In W. Chafe and J. Nichols (eds.), *Evidentiality: The linguistic coding of epistemology*. Norwood, NJ: Ablex. pp. 261-272.

Chafe, W. (1985) Linguistic Differences produced by Differences between Speaking and Writing. In D. R. Olson, N. Torrance and A. Hildyard (eds.), *Literacy, Language, and Learning*. Cambridge: Cambridge University Press. pp. 83-113.

Cheshire, J. and Trudgill, P. (1998) *The Sociolinguistics Reader. Vol. 2. Gender and discourse*. London: Arnold.

Connolly, L. (2004) The Limits of Irish Studies: Historicism, Culturalism and Paternalism. *Irish Studies Review* 12 (2): 139-62.

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

Corrigan, K. P. (1992) I gcuntas De muin Bearla do na leanbhain': eisimirce agus an Ghaeilge sa naou aois deag. In P. O'Sullivan (ed.), *The Irish World Wide*. Leicester: Leicester University Press. pp. 143–161.

Cressy, D (1980) *Literacy and Social Order: Reading and Writing in Tudor and Stuart England*. Cambridge: Cambridge University Press.

DeHaan, K. A. (2010) Negotiating the Transnational Moment: Immigrant Letters as Performance of a Diasporic Identity. *National Identities* 12: 107-131.

Daybell, J. and Hinds, P. (2010) *Material Readings of Early Modern Culture: Texts and Social Practices, 1580-1730*. Basingstoke: Palgrave Macmillan.

De Beaugrande, R. (1984) *Text production: toward a science of composition*. Norwood, New Jersey: Ablex Publishing Corporation.

De Fina, A. and King, K. A. (2011) Language problem or language conflict? Narratives of immigrant women's experiences in the US. *Discourse Studies* 13: 163-188.

Dossena, M. (2010) ‘A steedy hand, a geasent throat, a dry heart and an empty pip’: Scots and vernacular features in William Cameron’s letters. In R. M. Millar (ed.), *Northern Lights, Northern Words. Selected Papers from the FRLSU Conference*. Aberdeen: Forum for Research on the Languages of Scotland and Ireland. pp. 122-42.

Dossena, M. ‘Many strange and peculiar affairs’: Description, Narration and Evaluation in Scottish Emigrants’ Letters of the Nineteenth Century. *Scottish Language* 27 (2008).

Dossena, M. (2007) ‘As this leaves me at present’ – Formulaic usage, politeness and social proximity in nineteenth-century Scottish emigrants’ letters. In Elspaβ, S., Langer, N., Scharloth, J. and Vandenbussche, W. (eds.), *Germanic Language Histories from Below (1700-2000)*. Berlin: De Gruyter. pp. 13-29.

Dossena, M. (2004) Towards a Corpus of Nineteenth-century Scottish Correspondence. *Linguistica e Filologia* 18: 195-214.

Elliott, B. S., Gerber, D. A. and Sinke, S. (eds.) (2006) *Letters Across Borders : Epistolary Practices of International Migrants*. London: Palgrave Macmillan.

Elspaβ, S. (2007a) Everyday language in emigrant letters and its implications on language historiography – the German case. *Multilingua*. 26 (2/3): 151-165.

Elspaβ, S. (2007b) A twofold view ‘from below’: New perspectives on language histories and historical grammar. In S. Elspaβ, N. Langer, J. Scharloth and W. Vandenbussche (eds.), *Germanic Language Histories ‘from Below’ (1700-2000)*. Berlin: Walter de Gruyter. pp. 3-12.

Elspaβ, S., Langer, N., Scharloth, J. and Vandenbussche, W. (eds.) (2007) *Germanic language histories ‘from below’ (1700-2000)*. Berlin: Mouton de Gruyter.

Elspaβ, S. (2002) Standard German in the 19th-century? (Counter-) Evidence from the private correspondence of ‘ordinary people’. In A. R. Linn and N. McLelland (eds.), *Standardization: Studies from the Germanic Languages*. Amsterdam: Benjamins. pp. 43-65.

Emmons, D. M. (1990) *The Butte Irish: Class and Ethnicity in an American Mining Town, 1875-1925*. Urbana: University of Illinois Press.

Erickson, C. (1972) *Invisible Immigrants: The Adaptation of English and Scottish Immigrants in Nineteenth-Centrury America*. London: Weidenfeld & Nicolson.

Eyford, R. C. (2015) ‘Close together, though miles apart’: Family, Distance, and Emotion in the Letters of the Taylor Sisters, 1881-1921. *Histoire sociale/Social History* 47 (96): 67-86. [DOI: 10.1353/his.2015.0012]

Jeffries, L. and McIntrye, D. (2010) *Stylistics*. Cambridge: Cambridge University Press.

Fairman, T (2015) Language in print and handwriting. In A. Auer, D. Schreier and R. J. Watts (eds.), *Letter Writing and Language Change.* Cambridge, Cambridge University Press. pp 53-71.

Fairman, T (2012) Letters in mechanically-schooled language. In M. Dossena and G. Del Lungo Camiciotti (eds.), *Letter Writing in Late Modern Europe*. Amsterdam: John Benjamins Publishing Company. pp. 205-227.

Fairman, T. (2009) She has four and big agane: ellipses and prostheses in mechanically-schooled writing in England, 1795–1834. In I. Tieken-Boon van Ostade and W. van der Wurf (eds.), *Current Issues in Late Modern English*. Bern: Peter Lang. pp. 409-429.

Fairman (2008) The Study of Writing in Sociolinguistics. In S. Kermas and M. Gotti (eds.), *Socially-Conditioned Language Change: Diachronic and Synchronic Insights*. Lecce: Edizioni del Grifo. pp. 53-75.

Fairman, T. (2000) English pauper letters 1800-34 and the English language. In D. Barton and N. Hall (eds.), *Letter writing as a social practice*. Amsterdam: John Benjamins. pp. 63-82.

Fay, C. R. (1951) *Huskisson and his Age*. London: Longmans.

Fitzmaurice, S. (2007) Questions of standardization and representativeness in the development of social networks based corpora: the story of the Network of Eighteenth-century English Texts. In J. C. Beal, K. Corrigan, H. Moisl (eds.), *Creating and Digitizing Unconventional Corpora. Volume 2: Diachronic Corpora*. London: Palgrave. pp. 49-81.

Fitzpatrick, D. (1994) *Oceans of Consolation: Personal Accounts of Irish Migration to Australia*. Cork: Cork University Press.

García-Bermejo Giner, M. F. and Montgomery, M. (1997) British regional English in the nineteenth century: the evidence from emigrant letters. In A. R. Thomas (ed.), *Issues and Methods in Dialectology*. Bangor, Wales: University of Wales. pp. 167-183.

Gee, J. (2008) *Social linguistics and literacies: ideology in discourse*. London: Taylor & Francis.

Gerber, D. A. (2006) *Authors of Their Lives: The Personal Correspondence of British Immigrants to North America in the Nineteenth Century*. New York: New York University Press.

Griffin, G. (2006) Interdisciplinarity in Interdisciplinary Research Programmes in the UK. *Research Integration*. Hull: University of Hull. Available from: http://www.hull.ac.uk/researchintegration.

Halliday, M. A. K. and Matthiessen, C. M. I. M. (2004) *An Introduction to Functional Grammar*, Third Edition. London: Arnold.

Harper, M. (2010) Opportunity and Exile: Snapshots of Scottish Emigration to Australia. *Australian Studies* 2 (2): 1-21.

Helbich W. and Kamphoefner, W. D. (2006) How Representative are Emigrant Letters? An Exploration of the German Case. In B. S. Elliott, D. A. Gerber and S. M. Sinke (eds.), *Letters across Borders: The Epistolary Practices of International Migrants*. New York: Palgrave Macmillan. pp. 29-55.

Hitchcock, T. (2004) A new history from below. *History Workshop Journal.* 57 (1): 294–299.

Hitchcock, T., King, P. and Sharpe, P. (1996) *Chronicling Poverty The Voices and Strategies of the English Poor, 1640-1840*. London: Palgrave Macmillan.

Hoey, M. (2005) *Lexical Priming: A new theory of words and language*. Oxon: Routledge.

Huddleston, R. and Pullum, G. (2002) *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hunston, S. (2006) Phraseology and system: a contribution to the debate. In S. Hunston and G. Thompson (eds.), *System and corpus: exploring connections*. London: Equinox. pp. 55-80.

Hunston, S. (2002) *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hunter, M. (2009) *Editing Early Modern Texts. An Introduction to Principles and Practice*. Basingstoke: Palgrave Macmillan.

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford Univerity Press.

Hoey, M. (2001) *Textual Interaction: An introduction to written discourse analysis*. London: Routledge.

Hoey, M. (2005) *Lexical Priming: A new theory of words and language*. London: Routledge.

Holmes, J. (1995) *Women, men and politeness*. London: Longman.

Kamphoefner, W. D., Helbich, W. and Sommer, U. (1988) *News from the Land of Freedom: German Immigrants Write Home*. Ithaca: Cornell University Press.

Kaye, H. (1984) *The British Marxist Historians: An Introductory Analysis*. Cambridge: Polity.

Kindberg, T. and Hawke, S. (eds.) *RFC 41512005 IETF Tools: The 'tag' URI Scheme*. Available from: https://tools.ietf.org/html/rfc4151.

Koch, P. and Oesterreicher, W. (1985) Sprache der Nähe – Sprache der Distanz. Mündlichkeit and Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*. 36: 15-42.

Koch, P. and Oesterreicher, W. (1994) Schriftlichkeit und Sprache. In H. Günther and O. Ludwig (eds.), *Writing and Its Use. An interdisciplinary Handbook of International Research, vol. 1*. Berlin: Mouton de Gruyter. pp. 587-604.

Koos, G. (2001) The Irish Hedge Schoolmaster in the American Backcountry. *New Hibernia Review* 5: 9-26.

Kotthoff, H. and Wodak, R. (1997) *Communicating gender in context*. Amsterdam: Benjamins.

Kyotö, M., Grund, P. and Walker T. (2007) Regional variation and the language of English witness depositions 1560-1760: constructing a 'linguistic' edition in electronic form. In P. Pahta, I. Taavitsainen, T. Nevalainen and J. Tyrkkö (eds.), *Towards Multimedia in Corpus Studies (Studies in Variation, Contacts and Change in English 2)*. Helsinki: Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. Available from: http://www.helsinki.fi/varieng/series/volumes/02/kyto_et_al/.

Lass, R. (2004) 'Ut custodian litteras: Editions, Corpora and Witnesshood. In M. Dossena and R. Lass (eds.), *Methods and Data in Historical Dialectology. Linguistic Insights 16*. Bern: Peter Lang. pp. 21-48.

Levinson, S. C. (1983) *Pragmatics*. Cambridge: Cambridge University Press.

Lodge, D. (2002) *Consciousness and the Novel: Connected Essays*. Cambridge, Massachusetts:

Loomba, A. (1998) *Colonialism/Postcolonialism*. London: Routledge.

Lyons, M. (2010) A New History from Below? The Writing Culture of European Peasants, c. 1850 – c. 1920. *History Australia* 7 (3): 59.1-59.9. [DOI: 10.2104/ha100059]

López-Couso, M. J. and Méndez-Naya, B. (2010) Looks like, seems like, sounds like: Emerging evidential markers? [Paper presented at *ICAME31*, University of Giessen, 26–30 May].

López-Couso, M. J. and Méndez-Naya, B. (2011) The construction is quite old, it seems: Origin and development of the evidential/epistemic parenthetical 'it seems'. [Paper presented at the Helsinki Corpus Festival: The past, present and future of English historical corpora, University of Helsinki, 28 September – 2 October].

López-Schmidt, K. J. and Poulser, B. (2002) *Writing Peasants. Studies on Peasant Literacy in Early Modern Northern Europe*. Gylling: Landbohistorisk selskab.

Martínez, L. (2006) 'Cartas Migrantes': La correspondencia de una familia de asturianos en Chile (1874-1932). Unpublished preliminary thesis. Universidad de Alcalá de Henares.

McCarthy, A. (2005) *Irish Migrants in New Zealand, 1840-1937: 'the Desired Haven'*. Suffolk: The Boydell Press.

McLelland, N. (2007) 'Doch mein Mann möchte doch mal wissen…' A discourse analysis of 19th-century emigrant men and women's private correspondence. In S. Elspaβ, N. Langer, J. Scharloth and W. Vandenbussche (eds.), *Germanic Language Histories from Below (1700-2000)*. Berlin: De Gruyter. pp. 45-68.

Miller, K. A. (2008) *Ireland and Irish America: Culture, Class, and Transatlantic Migration*. Dublin: Field Day

Miller, K. A. (1985) *Emigrants and Exiles: Ireland and the Irish Exodus to North America*. New York: Oxford University Press.

Miller, K. A., Doyle, D.N. and Kelleher, P. (1995) For Love and Liberty: Irish Women, Migration and Domesticity in Ireland and America, 1815–1920. In P. O'Sullivan (ed.), *The Irish World Wide*. Leicester: Leicester University Press. pp. 54–61.

Miller, K. A., Schrier, A., Boling, B. D. and Doyle, D. N. (2003) *Irish Immigrants in the Land of Canaan: Letters and Memoirs from Colonial and Revolutionary America, 1675-1815*. New York: Oxford University Press.

Millett, Bella (2013) Whatever happened to electronic editing? In, V. Gillespie and A. Hudson, A. (eds.), *Probable Truth: Editing Medieval Texts from Britain in the Twenty-First Century.* Turnhout: Brepols. pp. 39-54.

Milroy, J. (2012) Sociolinguistics and Ideologies in Language History. In J. M. Hernández-Campoy and J. C. Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, First Edition. Chichester: Wiley-Blackwell. pp. 571-584.

Montgomery, M. B. (1995) The linguistic value of Ulster emigrant letters. *Ulster folklife* 41: 26-41.

Moreton, E. (2012) Profiling the Female Emigrant: A Method of Linguistic Inquiry for Examining Correspondence Collections. *Gender & History* 24 (3): 617-646.

Moreton, E. (2015, *forth*) 'I hope you will write': The function of projection structures in a corpus of nineteenth century Irish emigrant correspondence. *Journal of Historical Pragmatics* 16 (2).

Moreton, E., O'Leary, N. and O'Sullivan, P. (2014) Visualising the emigrant letter. *Revue Européenne des Migrations Internationales, Special issue: Traces of Dispersion*. 30 (3 & 4) 49-69. [http://remi.revues.org/?lang=en].

Nevalainen, T. and Raumolin-Brunberg, H. (1996) The Corpus of Early English Correspondence. In T. Nevalainen and H. Raumolin-Brunberg (eds.), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi. pp. 39-54.

Nolan, J. A. (1989) *Ourselves Alone: Women's Emigration from Ireland, 1885-1920*. Lexington: University Press of Kentucky.

Noonan, A. J. M. (2011) 'Oh those long months without a word from home', Migrant letters from mining frontiers. Cork: The Boolean, University College Cork. pp. 129–135. [Available from: http://publish.ucc.ie/boolean/].

Norlund, T. (2007) Double diglossia – lower class writing in 19th century Finland. *Multilingua* 22 (2/3): 229-246.

Nurmi, A. and Palander-Collin, M. (2008) Letters as a Text Type: Interacting in Writing. In M. Dossena and I. Tieken-Boon van Ostade (eds.), *Studies in Late Modern English Correspondence: Methodology and Data*. Bern: Peter Lang. pp. 21-49.

O'Farrell, P. (1984) *Letters from Australia 1825-1929*. New South Wales: New South Wales University Press.

O'Farrell, P. (1987) *The Irish in Australia*. New South Wales: New South Wales University Press.

O'Farrell, P. (1990) *Vanished Kingdoms, Irish in Australia and New Zealand, A Personal Excursion*. New South Wales: New South Wales University Press.

O'Hanlon, R. (1988) Recovering the Subject Subaltern Studies and Histories of Resistance in Colonial South Asia. *Modern Asian Studies* 22 (1): 189-224.

O'Sullivan, P. (2003) Developing Irish Diaspora Studies: A Personal View. *New Hibernia Review* 7 (1): 130-148.

Palander-Collin, M. (2009) Patterns of interaction: Self-mention and addressee inclusion in letters of Nathaniel Bacon and his correspondents. In A. Nurmi, M. Nevala and M. Palander-Collin (eds.), *The Language of Daily Life in England (1400–1800)*. Amsterdam and Philadelphia: John Benjamins Publishing Company. pp. 53-74.

Palander-Collin, M. (1999) Male and female styles in 17th century correspondence: I THINK. *Language Variation and Change* 11: 123–141.

Pine, C. (2003-2014) *Learn to Program*. Available from: https://pine.fm/LearnToProgram/.

Plummer, K. (2001) *Documents of Life 2: An Invitation to a Critical Humanism*. London: SAGE Publications.

Poland, B. and Pederson, A. (1998) Reading Between the Lines: Interpreting Silences in Qualitative Research. *Qualitative Inquiry* 4: 293-312.

Quirk, R., Greenbaum, S., Leech G. and Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Rayson, P. (2008) From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13 (4): 519-549. [DOI: 10.1075/ijcl.13.4.06ray].

Richards, E. S. (2010) Australian Colonial Mentalities in Emigrant Letters. *Australian Studies* 2: 1-17.

Richards, E. S. (2006) The limits of the Australian emigrant letter. In B. S. Elliot, D. A. Gerber and S. M. Sinke (eds.), *Letters across Borders: The Epistolary Practices of International Migrants*. New York: Palgrave Macmillan. pp. 56-74.

Säily, T., Nevalainen, T. and Siirtola, H. (2011) Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26 (2): 167–185.

Sairio, A. (2009) *Language and Letters of the Bluestocking Network. Sociolinguistic Issues in Eighteenth-century Epistolary English* (Monograph). (Mémoires de la Société Néophilologique de Helsinki 75). Helsinki: Société Néophilologique.

Sairio, A. (2005) 'Sam of Streatham Park': A linguistic study of Dr. Johnson's membership in the Thrale family. *European Journal of English Studies* 9 (1): 21–35.

Schrier, A. (1958) *Ireland and the Irish Emigration, 1850-1900*. Minneapolis: University of Minnesota Press.

Schuchard, R. (2002) Excavating the Imagination: Archival Research and the Digital Revolution. *Libraries & Culture* 37: 57-63.

Scott, J. W. (1992) Experience. In J. Butler and J. W. Scott (eds.), *Feminists Theorize the Political*. New York and London: Routledge. pp. 22-40.

Scott, M. and Tribble, C. (2006) *Textual patterns: key words and corpus analysis in language education*. Amsterdam: John Benjamins Publishing Company.

Seifert, S., Illetschko, M. and Stadler, P. *Towards a correspondence module in the TEI*. [Presentation at the TEI Conference, Evanston, 22 October 2014.] Available from: https://github.com/TEI-Correspondence-SIG/correspDesc/tree/master/presentations.

Sifton, P. G. (1977) The Provenance of the Thomas Jefferson Papers. *American Archivist* 40 (1): 17–30.

Simpson, P. (1993) *Language, Ideology and Point of View*. London: Routledge.

Sokoll, T. (2001) *Essex Pauper Letters, 1731-1837*. Oxford: Oxford University Press for the British Academy.

Spivak, G. C. (2006) Can the Subaltern Speak? In B. Ashcroft, G. Griffiths and H. Tiffin (eds.), *The Post-Colonial Studies Reader*, Second Edition. Oxford: Routledge. pp. 28-37.

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

Steen, S. J. (1994) Manuscript Matters: Reading the Letters of Lady Arbella Stuart. *South Central Review*. 11: 24-38.

Stubbs, M. (2008) Conrad in the computer: examples of quantitative stylistic methods. In R. Carter and P. Stockwell (eds.), *The Language and Literature* Reader. London: Routledge. pp. 230-243.

Talbot, M. (1998) *Language and gender*: *An introduction*. Oxford: Blackwell.

TEI Consortium (eds.), *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. [Version: 2.8.0]. [Last modified: 2015-05-06]. TEI Consortium. Available from: http://www.tei-c.org/Guidelines/P5/.

TEI SIG: Correspondence. Available from: http://wiki.tei-c.org/index.php/SIG:Correspondence#Introduction.

Thomas, W. I. and Znaniecki, F. (1919-1920) *The Polish Peasant in America*, 5 volumes (Vols. 1 & 2, 1918; Vols. 3, 4 & 5, 1919-1920). New York: Dover Publications Inc.

Thompson, G. (2012) Intersubjectivity in newspaper editorials. *English Text Construction. Special issue: Intersections of Intersubjectivity* 5 (1): 77–100.

Thompson, G. and Thetela, P. (1995) The sound of one hand clapping: The management of interaction in written discoures. *Text & Talk* 15 (1): 103–27.

Thompson, S. A. and Mulac, A. (1991) A quantitative perspective on the grammaticization of epistemic parentheticals in English. In E. C. Traugott and B. Heine (eds.), *Approaches to Grammaticalization*. Amsterdam and Philadelphia: John Benjamins. pp. 313–329.

Tieken-Boon van Ostade, I. M. (2010) *The Bishop's Grammar: Robert Lowth and the Rise of Prescriptivism*. Oxford: Oxford University Press.

Toolan, M. (2009) *Narrative Progression in the Short Story*. Amsterdam: John Benjamins Publishing Company.

Traugott, E. C. and Dasher, R. B. (2002) *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Van Dijk, T. A. (1977) *Text and Context: Explorations in the semantics and pragmatics of discourse*. London: Routledge.

Van Leeuwen, M. H. D. and Maas, I. (2011) *HISCLASS. A historical international social class scheme*. Leuven: Leuven University Press.

van der Vaal, M., Rutten, G. and Simons, T. (2012) Letters as loot: Confiscated letters filling major gaps in the history of Dutch. In M. Dossena and G. Del Lungo Camiciotti (eds.), *Letter Writing in Late Modern Europe*. Amsterdam: John Benjamins Publishing Company. pp. 139-161.

Vandenbussche, W. (2006) A Rough Guide to German Research on 'Arbeitersprache' during the 19th Century. In H. Andrásová, P. Ernst and L. Spácilová. *Germanistik genießen: Gedenkschrift für Doc. Dr. phil. Hildegard Boková*. Wien: Praesens Verlag. pp. 439-458.

Vanhoutte, E. and Van den Branden, R. (2009) Describing, transcribing, encoding, and editing modern correspondence material: a textbase approach. *Literary and Linguistic Computing* 24 (1): 77-98.

Watt, R. J. (2012) Language Myths. In J. M. Hernández-Campoy and J. C. Conde-Silvestre (eds.), *The Handbook of Historical Sociolinguistics*, First Edition. Chichester: Wiley-Blackwell. pp. 585-606.

Watts, R. and Trudgill, P. (2002) *Alternative Histories of English*. London: Routledge.

Wei-Ling Wee, C. (2009) *Mobilising Action through Management Email Texts: The Negotiation of Evaluative Stance through Choices in Grammar and Disourse* (PhD dissertation). Department of Languages and Linguistics, Faculty of Arts & Social Sciences, University of New South Wales.

Wodak, R. (1997) *Gender and discourse*. London: Sage Publications.

Yokoyama, O. T. (2008) *Russian Peasant Letters*. Wiesbaden: Harrassowitz.

Zamper, S. (1991) *In Their Own Words: Letters from Norwegian Immigrants*. Minneapolis: University of Minnesota Press.

Appendix A

Differences between the markup proposed in chapter five (example A) and the TEI correspondence SIG's current module (example B).

Example A

```
<profileDesc>
    <ct:correspDesc xmlns="http://wiki.tei-c.org/index.php/SIG:Correspondence/task-
    force-correspDesc">
        <ct:participant role="sender"
            <persName key="LOUGH_Pers0001">Elizabeth Lough</persName>
            <placeName key="LOUGH_Place0001">Winsted, Connecticut</placeName>
            <date when="1876-03-07"/>
        </ct:participant>
        <ct:participant role="recipient"
              <persName key="LOUGH_Pers0007">Elizabeth McDonald Lough</persName>
              <persName key="LOUGH_Pers0008">James Lough</persName>
              <persName key="LOUGH_Pers0002"/>Alice Lough</persName>
              <persName key="LOUGH_Pers0003"/>Anne Lough</persName>
              <persName key="LOUGH_Pers0004"/>Julia Lough</persName>
              <persName key="LOUGH_Pers0005"/>Mary Lough</persName>
              <placeName key="LOUGH_Place0006"/>Meelick, Queen's County</placeName>
        </ct:participant>
    </ct:correspDesc>
</profileDesc>
```

Example B

```
<profileDesc>
    <correspDesc>
        <correspAction type="sending">
            <persName key="LOUGH_Pers0001">Elizabeth Lough</persName>
            <settlement> key="LOUGH_Place0001">Winsted, Connecticut</settlement>
            <date when="1876-03-07"/>
        </correspAction>
        <correspAction type="receiving">
            <persName key="LOUGH_Pers0007">Elizabeth McDonald
Lough</persName>
            <persName key="LOUGH_Pers0008">James Lough</persName>
            <persName key="LOUGH_Pers0002"/>Alice Lough</persName>
            <persName key="LOUGH_Pers0003"/>Anne Lough</persName>
            <persName key="LOUGH_Pers0004"/>Julia Lough</persName>
            <persName key="LOUGH_Pers0005"/>Mary Lough</persName>
            <settlement key="LOUGH_Place0006"/>Meelick, Queen's
County</settlement>
        </correspAction>
    </correspDesc>
</profileDesc>
```

Note: while the markup detailed in chapter five uses the <participant> element and @role attribute to distinguish between "sender" and "recipient", the current TEI correspondence SIG proposal uses the <correspAction> element and @type attribute to distinguish between the act of "sending" and "receiving". Additionally the <settlement> element is used instead of the <placeName> element to provide details of the sender/recipient's location.