Identifying speech acts in a corpus of historical migrant correspondence

Rachele De Felice, and Emma Moreton

Author post-print (accepted) deposited by Coventry University's Repository

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

DOI 10.1080/00393274.2019.1616216
ISSN 0039-3274
ESSN 1651-2308

Publisher: Taylor and Francis

This is an Accepted Manuscript of an article published by Taylor & Francis in Studia Neophilologica on 20th June 2019, available online:
http://www.tandfonline.com/10.1080/00393274.2019.1616216

Copyright © and Moral Rights are retained by the author(s) and/ or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This item cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder(s). The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holders.

This document is the author's post-print version, incorporating any revisions agreed during the peer-review process. Some differences between the published version and this version may remain and you are advised to consult the published version if you wish to cite from it.

Identifying speech acts in a corpus of historical migrant correspondence

Rachele De Felice and Emma Moreton

Abstract

A full account of the pragmatics of personal correspondence requires speech act annotation, and as manual annotation of large dataset can be extremely difficult, this essay proposes to use an automated speech act tagger developed by De Felice et al. (2013). It was originally designed for use with business emails; however the latest iteration of the tagger can be applied to other datasets – such as personal correspondence – providing a useful resource for the corpus linguistics community. In this essay, the speech act tagger is tested on a collection of letters written by Irish migrants at the end of the nineteenth century. After discussing issues to do with the digitisation, transcription and annotation of historical migrant correspondence, the essay will report on the results of this trial study, demonstrating how the tagger can perform with some success even on corpora with very different characteristics. Although the dataset used for this trial study is small, the findings show the potential for carrying out this type of analysis across larger digital archives allowing for different datasets to be compared, taking into consideration sociobiographic variables such as the author's sex, class and role within the notional familial hierarchy.

1. Introduction

Over the past decade or so, 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 distant times. As argued by Nevalainen and Raumolin-Brunberg, 'correspondence often resembles spoken registers more closely than other types of writing', thereby providing linguists with a window onto how language was used at a particular period in time, as well as providing information about the letter writers themselves and the historical context in which they wrote (1996: 39, citing Biber 1995: 283–300).

A full account of the pragmatics of personal correspondence requires speech act annotation, whereby each utterance in the corpus is assigned a category such as 'request', 'commitment', or 'expression of feeling'. This enables the user to analyse the data in an inclusive manner, considering all examples of a given speech act regardless of their surface form. For large datasets it would be extremely difficult – as well as time-consuming and costly – to carry out the annotation manually. This essay proposes to use instead an automated speech act tagger developed by De Felice et al. (2013).

The speech act tagger (henceforth SAT) is tested on a small corpus of nineteenth century Irish migrant letters (the Lough letters). After discussing issues concerning the digitisation, transcription and annotation of historical migrant correspondence, the essay will report on the results of this trial study, demonstrating how the tagger can perform with some success even on corpora with very different characteristics from those for which it was originally designed. Although the dataset used for this trial study is small (just one letter series containing 39 texts), the findings show the potential for carrying out this type of analysis across larger digital archives, thereby allowing different datasets to be compared.

2. The Lough letters

The Lough (pronounced 'lɒk') family letters are from Professor Kerby Miller's collection of Irish migrant correspondence, held at the University of Missouri.[1] The collection contains letters by four sisters who migrated from Ireland to America in the 1870s and 1880s (an additional two sisters remained in Ireland). Significantly, these letters are drawn from a much larger body of Irish migrant correspondence collected by

1 Professor Kerby Miller, Emeritus Professor, Department of History at the University of Missouri: https://en.wikipedia.org/wiki/Kerby_A._Miller.

Miller. Miller himself has explored this wider corpus in several pioneering works on Irish migration (see, for instance, Miller 1985 and Miller et al. 2003) and his archive of over 5,000 letters has been referred to by many scholars including Emmons (1990), Corrigan (1992), Koos (2001), Bruce (2006) and Noonan (2011).

In the early 1950s, some 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, later Professor Emeritus at the University of Cincinnati, who subsequently used 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 analysed in most detail. In his 2008 study, *Ireland and Irish America: Culture, Class, and Transatlantic Migration*, Miller 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: 307).

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 from Ireland to America. 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 impartible inheritance systems (in turn, leading to changes in marriage trends), contributed to 'a massive post-famine emigration by young, unmarried women' (Miller 1985: 3). By the second half of the nineteenth century Ireland had become 'a nation characterized by late and reluctant marriage as well as by a massive voluntary exodus' (ibid.: 8). Between 1852 and 1921 the median age for female Irish migrants was 21.2 and after 1880 young women, such as the Lough sisters, constituted the majority of the departing Irish (Miller 1985: 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 six Lough sisters – Elizabeth, Alice, Annie, Julia, Mary and Maggie – came from a Roman Catholic family in Meelick, in what was then called Queen's County (now County Laois), Ireland. The six sisters were daughters of Elizabeth McDonald Lough and James Lough who lived on a smallholding consisting of two fields, one of which, according to family legend, was sold to pay for the sisters' passages. The Lough family was, according to Miller, not of the lowest class as both parents and daughters were able to write. Apart from Mary and Maggie, all the Lough sisters migrated to America between 1870 and 1884. The sisters who migrated were, in Miller's words, four 'very dutiful, hard-working, and pious Irish females'.[2] 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).

This essay focuses on the letters of just one of the Lough sisters – Annie (sometimes referred to in the correspondence as Nan or Nannie). Annie was the third sister to migrate in 1878, following her older sisters Elizabeth and Alice. She lived in Winsted, Litchfield County, Connecticut all her life, where she appears to have worked as a servant for a while. Annie married John McMahon on 9 June 1886 – a labourer or factory worker – however, she bore no children. Annie died in Winsted in 1935; her husband died on 18 September 1936.

Table 1 shows how frequently Annie wrote home to Ireland and to whom her letters were addressed. There are 39 letters by Annie in the Lough files. Annie's earlier letters were addressed to her mother. The first of these was sent in around 1878 (although the letter itself is not dated) from Queenstown, County Cork, Ireland, just before Annie set sail for America. After 1895 (around the time of her mother's death), Annie starts writing to her sister, Mary, and the correspondence continues into the late 1920s. Annie writes to Mary regularly during this 30 to 35-year period, often sending letters at Easter and Christmas, or on the

2 This quotation is taken from Miller's notes in the Lough file.

anniversary of a family member's death. Annie's letters are fairly evenly distributed and there are no major gaps in her correspondence. In the 1910s-20s Annie writes to her two nieces (Kate and Alice) and her nephew (James). The content of these letters suggests that Annie maintained regular contact with her nieces and nephew in Ireland; however, we do not have copies of these other letters. The letters have been sorted chronologically in Table 1; although some of the letters are not dated their content would suggest they were written from 1920 onwards. Information about the location of the sender and recipient is also provided.

Table 1. Annie Lough collection

Ref.	Day	Month	Year	From (location)	Recipient	To (location)
1	18	June	-	Queenstown, Ireland	Mother	Meelick, Ireland
2	03	March	1890	Winsted, America	Mother & Sister	Meelick, Ireland
3	29	October	1891	Winsted, America	Mother	Meelick, Ireland
4	15	December	1891	Winsted, America	Mother	Meelick, Ireland
5	23	March	1892	Winsted, America	Mother	Meelick, Ireland
6	30	March	1893	Winsted, America	Mother	Meelick, Ireland
7	-	December	-	Winsted, America	Mother	Meelick, Ireland
8	-	-	-	Winsted, America	Mother	Meelick, Ireland
9	17	March	1895	Winsted, America	Sister	Meelick, Ireland
10	18	May	1899	Winsted, America	Sister	Meelick, Ireland
11	16	February	1901	Winsted, America	Sister	Meelick, Ireland
12	21	September	1901	Winsted, America	Sister	Meelick, Ireland
13	10	December	1902	Winsted, America	Sister	Meelick, Ireland
14	03	April	1906	Winsted, America	Sister	Meelick, Ireland
15	20	June	1906	Winsted, America	Sister	Meelick, Ireland
16	30	November	1906	Winsted, America	Sister	Meelick, Ireland
17	12	December	1912	Winsted, America	Sister	Meelick, Ireland
18	08	December	1913	Winsted, America	Sister	Meelick, Ireland
19	11	December	1914	Winsted, America	Niece	Meelick, Ireland
20	31	April	1918	Winsted, America	Sister	Meelick, Ireland
21	06	May	1918	Winsted, America	Sister	Meelick, Ireland
22	14	July	1918	Winsted, America	Sister	Meelick, Ireland
23	14	August	1919	Winsted, America	Sister	Meelick, Ireland
24	21	March	1920	Winsted, America	Niece	Ireland
25	21	March	1920	Winsted, America	Sister	Meelick, Ireland
26	01	December	1919	Winsted, America	Sister	Meelick, Ireland
27	07	December	1919/1920	Winsted, America	Sister	Meelick, Ireland
28	-	-	-	Winsted, America	Sister	Meelick, Ireland
29	31	March	1924	Winsted, America	Sister	Meelick, Ireland
30	29	September	1925	Winsted, America	Sister	Meelick, Ireland
31	28	March	1928	Winsted, America	Sister	Meelick, Ireland
32	18	October	1928	Winsted, America	Sister	Meelick, Ireland
33	04	November	-	Winsted, America	Nephew	Ireland
34	-	-	-	Winsted, America	Sister	Meelick, Ireland
35	-	-	-	Winsted, America	Sister	Meelick, Ireland
36	-	-	-	Winsted, America	Sister	Meelick, Ireland
37	-	-	-	Winsted, America	Sister	Meelick, Ireland
38	-	-	-	Winsted, America	Sister	Meelick, Ireland
39	-	-	-	Winsted, America	Sister	Meelick, Ireland

3. Preparing the letters for analysis

In the Lough collection, in most cases, there is a photocopy of the original manuscript (Fig. 1) together with Miller's typed transcription (Fig. 2). Miller's transcriptions represent, as closely as possible, the language, structure and layout of the original manuscripts. That is to say, spelling variations and grammatical irregularities have not been standardised.

Figure 1. Photocopy of original manuscript

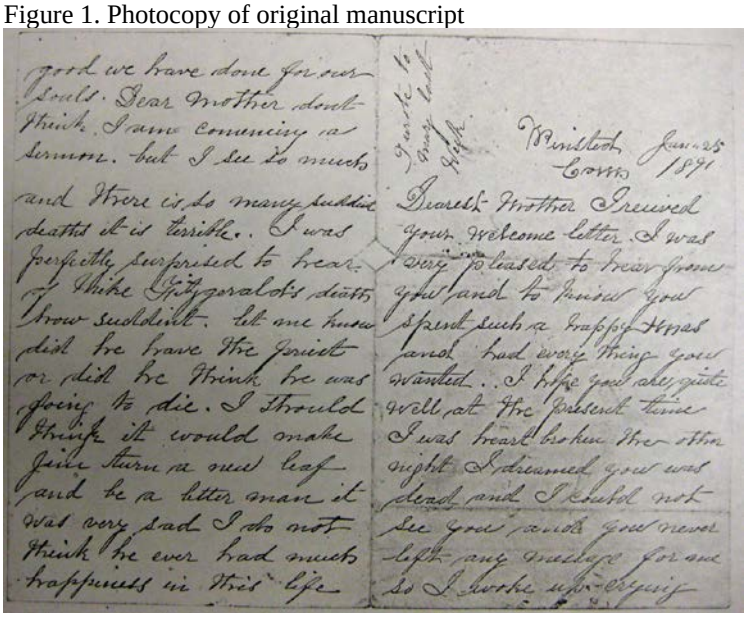
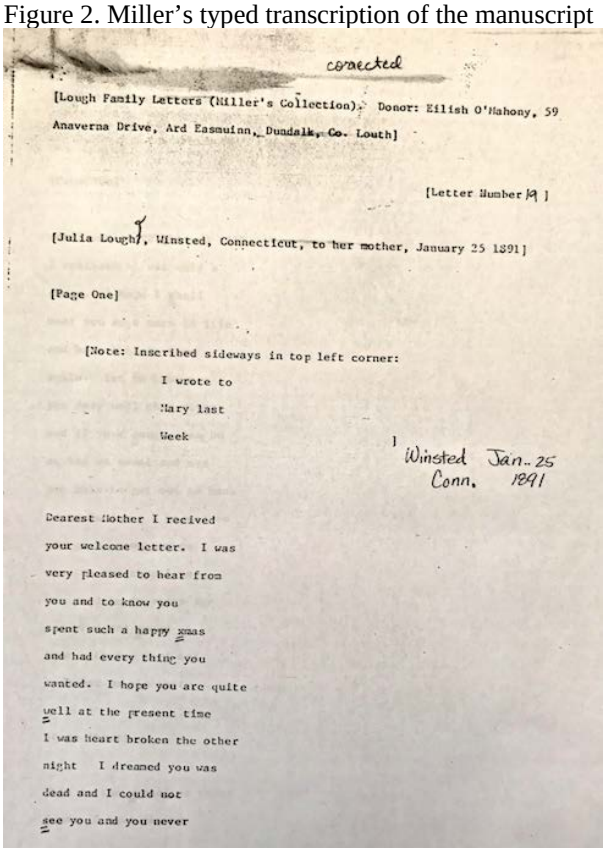


Figure 2. Miller's typed transcription of the manuscript



corrected

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

[Letter Number 19]

[Julia Lough, Winsted, Connecticut, to her mother, January 25 1891]

[Page One]

[Note: Inscribed sideways in top left corner:
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

Working with historical migrant letter collections poses various methodological challenges. Transcribing

letters with ‘inadequate paragraphing and punctuation, ungrammatical constructions [and] highly irregular spelling’ (Elliott et al. 2006: 4) – typical of many historical migrant letter corpora – can be particularly problematic. Indeed, whilst preserving the original spelling, structure and layout of the letter is crucial in some disciplines (historical and socio-linguistics, for instance), this can create issues when it comes to using computer software to analyse the texts. Tools such as *Sketch Engine* (Kilgarriff et al. 2004; Kilgarriff & Kosem 2012) or *Wmatrix* (Rayson 2009), for example, which automatically tag data for Part of Speech and, in the case of *Wmatrix*, semantic domains, require that spelling and grammatical variations be standardised for the taggers to work effectively. Similarly, the SAT (which will be described in detail in Section 4) is designed to be used with well-formed text. Variable spelling can cause problems for Part of Speech identification, as noted above, and by obscuring the presence of particular words and phrases used in speech act categorisation. The tagger also relies on the text being segmented into individual, clearly delimited sentences, as it operates on a sentence-by-sentence basis.

To prepare the Lough letters for analysis using the SAT, they first needed to be transcribed and saved in a digital format such as Plain Text or XML. As Annie’s letters contain very little punctuation, just 47 full-stops and 77 commas in all 39 letters, full-stops were inserted manually using the tag <punct type='stop'/>. Our approach, here, was 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 Annie’s letters. For example, although Annie rarely punctuates her writing, the occasional full-stops that she does use tend to indicate a change in topic (rather than the end of a sentence). Additionally, she appears to use vocatives to indicate a shift in the direction of the discourse. In the example below, for instance, the vocative *Dear Mother* indicates a move from the formulaic greeting (a variation of *I hope this letter finds you well as this leaves me at present*) to the subject of health and illness, a full-stop was thus inserted as follows:

> My dear Mother and Sister Mary I received your letter in due time and glad to hear you are all well and thankful for your nice letter all friends are well hoping these few lines will find you and Mary in good health**<punct type='stop'/>** *Dear Mother* I suppose this is the month you dred most with your cough (Annie to her mother and sister, 3 March 1890)

While vocatives tend to indicate a change in topic, clauses such as *I hope*, *I suppose*, *I wish*, *I am sure*, *I am glad* etc. (i.e. clauses which typically introduce another, secondary clause) tend to mark the beginning of a new sentence. In the example below, full-stops were inserted before *I am sure* and *I hope*, as follows:

> Mrs Odlurn lived a good age**<punct type='stop'/>** *I am sure* she will be missed and her death bring lots of changes**<punct type='stop'/>** *I hope* John will be kept in the work (Annie to her sister, 18 May 1899)

Finally, the conjunction *and* is often used in place of a full-stop to link semantically related clauses. While these, often long, text sections read quite clearly, they do cause problems for the SAT as excessively long sentences cannot be processed by the syntactic parser component of the tagger, leading to it simply not categorising the sentence at all. In cases such as the example below, where there is a long text section with *and* being used to link a number of clauses, full-stops were inserted where this was felt to be appropriate, as follows:

> **<punct type='stop'/>** *and* I hope these few lines will find you in good health and Kate and all the families**<punct type='stop'/>** *and* I am so glad were all so kind to you in your trouble**<punct type='stop'/>** *and* I am sure you are getting along as good as you can but it is hard to bear but I hope God will assist you and grant you strength to bear your loss before this letter comes to you (Annie to her sister, 28 March 1928)

Ultimately, the process of inserting full-stops was subjective and in doing this we offer just one, personal reading of the texts. Whilst it was possible to automate some of this process (inserting full-stops before the vocatives *Dear Mother* or *Dear Mary*, for instance), the rest had to be carried out manually. Table 2 shows the number of full-stops that were contained in the original manuscripts (the ‘Original’ column) and the number of full-stops that were inserted manually (the ‘Added’ column). The ‘Total’ column shows the total number of full-stops per letter after the preprocessing work had been carried out. While Annie’s earlier

letters (ref. 1 to 16, sent between 1878 and 1912) contain very few full-stops, she begins to punctuate her writing more consistently from letter 18 (sent on 8 December 1913) onwards. One letter in particular (ref. 20 – a letter sent on 31 April 1914) contains significantly more full-stops than the others, but this is somewhat anomalous.

Table 2. Full-stops in Annie's letters before and after preprocessing work

Ref.	Word count	Original	Added	Total
1	356	0	27	27
2	480	1	26	27
3	1055	0	90	90
4	487	0	41	41
5	1017	1	80	81
6	971	0	65	65
7	208	0	17	17
8	645	1	48	49
9	612	0	49	49
10	541	2	40	42
11	394	1	36	37
12	441	0	40	40
13	365	0	36	36
14	431	0	35	35
15	332	0	32	32
16	302	0	26	26
17	632	1	54	55
18	514	3	32	35
19	398	5	17	22
20	664	10	35	45
21	884	2	40	42
22	863	2	48	50
23	857	0	26	26
24	469	2	31	33
25	649	3	50	53
26	503	3	41	44
27	396	1	21	22
28	435	0	36	36
29	538	1	46	47
30	237	0	17	17
31	206	0	15	15
32	870	1	73	74
33	513	2	42	44
34	261	0	23	23
35	400	1	29	30
36	489	2	40	42
37	476	2	39	41
38	207	0	13	13
39	307	0	20	20

In terms of dealing with spelling variations, tools such as *VARD* (a preprocessing tool designed to deal with spelling variations in historical texts) can certainly save time when working with larger collections.[3]

3 *VARD*. UCREL (University Centre for Computer Corpus Research on Language), Lancaster University. Available from: http://ucrel.lancs.ac.uk/vard/about/ [Accessed 1 March 2017]. See also Baron & Rayson (2008).

However, as our dataset is relatively small (20,405 tokens) we decided to carry out this process manually, inserting the tags <sic> (to annotate the original spelling) and <corr> (to annotate the standardised spelling), as follows: <sic>untill</sic> <corr>until</corr>. Table 3 shows the most frequent spelling irregularities that were identified in Annie's letters (organised alphabetically) together with their corrected forms. The 'Freq.' column shows how often they occurred across all 39 letters.

Table 3. Typical spelling irregularities in Annie's letters

Original spelling	Standardised	Freq.
aint	isn't	2
all right	alright	10
any one	anyone	13
any thing	anything	5
any where	anywhere	1
babey	baby	3
cant	can't	7
deed	dead	3
dont	don't	37
didnt	didn't	2
dred	dread	1
every one	everyone	4
evry / evryone	every / everyone	7
good by	goodbye	42
newes	news	14
prey/s	prays	5
reed	read	13
rosery	rosary	2
ther	there	4
untill	until	10
what ever	whatever	1
wont	won't	13

Fig. 3 gives an example of one of Annie's letters, annotated for full-stops and standardised spellings.

Figure 3. Annotated XML version of a letter by Annie to her mother (December n.d.)

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <teiHeader>
    <fileDesc/>
    <sourceDesc/>
    <profileDesc/>
  </teiHeader>
  <body>
    <p>Dear Mother I hope you will enjoy yourself this Christmas and that you will spend a very
happy one<punct type='stop'/> you have all you want now to make you comfortable <sic>an</sic>
<corr>and</corr> I hope God bless you all and send you a merry Christmas and a happy new
year<punct type='stop'/> I am writing one sheet in this letter to Mary<punct type='stop'/> Dear
Mother all friends here are well<punct type='stop'/> Julie wrote to you this week<punct
type='stop'/> Alice has another young son<punct type='stop'/> it was not christened when she
wrote so I cannot tell you what she called him<punct type='stop'/> they are all well<punct
type='stop'/> Mary said she wrote to her but she did not answer it<punct type='stop'/> she has a
big family to work for and she does not have much time for writing<punct type='stop'/> Dear
Mother I hope you will write soon again and let me know all <sic>pirtuclers</sic>
<corr>particulars</corr><punct type='stop'/> and we all join in wishing you and Mary and John a
merry Christmas<punct type='stop'/> I hope Maggie is doing well for herself in London<punct
type='stop'/> she is apt to come home next summer<punct type='stop'/> <sic>good by</sic>
```

<**corr**>goodbye</**corr**> dear Mother for a while<**punct type='stop'**/> with best and fond love to you and you may be sure I will preserve your letter for a long time<**punct type='stop'**/> from your ever loving Nannie<**punct type='stop'**/></**p**>
</**body**>
</**TEI**>

Once all 39 of Annie's letters had been digitised and standardised for spelling variations, with full-stops inserted, it was possible to run the data through the SAT.

4. The Speech Act Tagger

The SAT used to process the Lough letters is an automated natural language processing (NLP) tool designed to categorise utterances into one of seven speech act classes on the basis of their lexical and grammatical features. The tagger was designed primarily to process contemporary email language, to identify the use of different speech acts particularly in the workplace domain. Its application to such a different text type – both as a genre and as a point in time – poses interesting challenges to its performance. An element of convergence does remain, however: both emails and the Lough letters are forms of correspondence, albeit ones from different domains.

Table 4. Summary of speech act categories used in the SAT

Speech act	**Tag**	**Example**
Direct request	DR	Please send me the files.
1st person commitment	FPC	I will attend the meeting.
1st person expression of feeling	FPF	I am uncertain about the agenda.
1st person other	FPO	I am an employee of this company.
Other statements	OT	The meeting is at 8 tomorrow. You always work so hard.
Open question	QQ	What time is the meeting?
Question-request	QR	Could you send me the files?

The speech act categories used by the tagger are summarised in Table 4 above. A comprehensive explanation of the process that led to these seven categories can be found in De Felice et al. (2013). The two key principles underpinning the choice of these categories are that they should be non-domain-specific and closely related to traditional speech act categories. This is to enable the application of the tagger to a wide range of data (as demonstrated in the present article) and to facilitate comparisons with other speech act research. A short explanation is given here to assist the reader in interpreting the findings discussed in this paper. Requests (similar to traditional directives) are divided into 'direct' (2nd person statements such as *you must*… and imperatives) and 'question' (canonical indirect requests formulated as interrogatives) requests. This distinction is made to allow for more fine-grained analysis of different types of requests. First person statements are divided into three categories: commitments (= traditional commissives), expressions of feelings (similar to traditional expressives), and others with less clear illocutionary force. The 'first person other' category – as opposed to 'other statements' which are those with 2nd and 3rd person subjects and variable illocutionary force – is introduced to aid a better understanding of how individuals speak of themselves, by contrasting it with commitments and expressions of feeling. Finally, information-seeking interrogatives are introduced as a separate category, 'open questions', as it can be helpful to analyse them as a stand-alone speech act type.

The tagger is trained on 20,700 sentences of American English emails from the Enron corporation, which have been manually annotated for speech acts by two native speakers (see De Felice et al. 2013). Full technical details are available in De Felice and Deane (2012); briefly described, the principle driving the tagger is each sentence is represented by a collection of features describing its syntactic and lexical properties, and labelled with the corresponding speech act tag. A machine learning classifier is trained to associate particular combinations of features to a given speech act category. It then uses this information to assign a previously unseen utterance to one of the categories. These features include Part of Speech tags, syntactic information, punctuation, use of modal verbs, verb transitivity, word clusters, and 'formulaic

structures' – a tagger-specific feature which refers to a set of particular combinations of syntactic and lexical items that are strongly associated with a given speech act (see De Felice & Deane 2012 for further details). The system is fully automated. It relies on the C&C toolkit (Curran, Clark & Bos 2007), which consists of a set of applications including a morphological analyser (Minnen, Carroll & Pearce 2001), a Part of Speech tagger, a parser (Clark & Curran 2007), and a named entity recogniser (Curran & Clark 2003). These tools allow us to record the lexical and syntactic information which forms the feature set. A simplified representation of a sentence and its associated features is below:

Table 5: Example sentence with associated features

I hope James Hickey is very well also [1st person expression of feeling]

Feature	Value
Subject	Pronoun, first person
First word	I
Last word	Also
Modal	None
Formulaic structure	None
Sentence type	Declarative
Verb type	Simple present
Key bigrams	I hope

On like-for-like data – that is, test data from the Enron corpus – the tagger achieves around 75% accuracy, though with noticeable variation in performance across different categories. More recently, the tagger has also been tested on British English email data (De Felice & Moreton 2015), giving average accuracy of 80%.[4] Its application to 19th century correspondence, however, presents a different kind of test of its robustness. The technical challenges of working with transcribed letters have already been outlined. A further issue arises not from the age but the genre of the correspondence, namely the fact that they are of a personal rather than a professional nature: the linguistic features expected by the tagger might be more typically found in the professional genre and simply not present in this dataset.

5. Overview of results

Because there is no pre-existing gold standard – that is, there is no manually annotated version of the letters against which to compare the tagger's performance – it is impossible to say precisely how accurate it has been. We can read each letter sentence by sentence and discuss whether we agree with its judgement, which leaves us open to bias, as we might be unduly influenced by the tagger's choice. As well as the risk of bias, we are dealing with 1465 utterances in total in this dataset, so the process would be impractical and time-consuming. However, we have examined 19 letters, consisting in total of 621 sentences (42% of the data) to gain an initial understanding of the tagger's general performance on this new dataset. Within this sample, our judgement is that the tagger has correctly categorised 494 of the sentences, that is, 79.5% accuracy. This is comparable with its performance on workplace emails, showing that the tool is robust enough to be applied to other domains, and can be used with confidence in the exploration and analysis of correspondence corpora. In Section 3, we described some of the edits that were applied to the letters to mitigate the effects of their language variation, such as standardising spelling. In previous attempts at tagging 19th century letters, non-standard spelling emerged as a problem for the tagger, and the positive impact of this editing procedure is evident in the improved accuracy shown here.

Overall, different speech act categories are used with different frequencies by Annie, as shown in Table 6, though it must be remembered that these figures are derived automatically and include incorrectly categorised sentences.

Table 6. Overview of the speech act categories identified in the Annie Lough collection

4 This discrepancy in accuracy is in all likelihood due to the British data having a simpler, more formulaic structure, which favours the tagger's performance.

Speech act category	Identified overall	Identified correctly in sample
Other statements	625 – 43% of total	248 – 50% of sample
1st person statement	293 – 20% of total	90 – 18% of sample
1st person expressive	284 – 19% of total	91 – 18% of sample
1st person commitment	114 – 8% of total	26 – 5% of sample
Directive	116 – 8% of total	27 – 5% of sample
Open question	31 – 2% of total	12 – 4% of sample
Question-directive	2 – 0.01% of total	0

A closer analysis of the data reveals that the average accuracy of 79.5% masks a range of performance across the different categories, illustrated in Table 7, which reports figures based on the subset of 19 letters which we have checked manually. Accuracy here refers to the percentage of sentences assigned to a given category which actually belong to that category (in other words, the tagger's precision).

Table 7. Accuracy of the SAT in a subset of 19 letters

Speech act category	Accuracy
Other statements	89%
1st person expressive	83.5%
Open question	80%
1st person statement	72%
1st person commitment	62%
Directive	53%

As in workplace email data, 'Other' statements (2nd and 3rd person utterances) are by far the easiest for the tagger to identify correctly; in this dataset, open questions and expressives also score very highly.[5] The latter in particular is a very encouraging result because the analysis of this speech act is an important aspect of the study of personal letters. The low rate of success in identifying directives requires further investigation. It is possible that this is caused by the very different nature of directives in personal letters and in workplace emails, such that the features normally associated with this speech act simply do not appear in the Lough letters, and perhaps in personal letters more generally.

To better understand the performance of the tagger on a range of sentences, we analyse several passages from one letter (ref. 19 – a letter to Annie's niece, Alice, dated 11 December 1914) in close detail, illustrating both correctly and incorrectly identified sentences.

(1) Dear Alice I thought I would like to write a few lines to you
(2) I have not heard a word from anyone from home since last Easter

Both the first two sentences of the letter are tagged as First Person Statements, correctly in our view, as they are simply stating some facts in the first person.

(3) Mother wrote then and she had a letter from me about the same time and she said she was going to write to me soon again but I think she forgets all about me but I hope Mother and Father and you and sister are all very well and I hope Sister Maggie and Husband and James is very well. all friends here are very well

This sentence has been tagged as 'Other' – a generic 2nd or 3rd person statement. This is certainly true of some of the clauses in this very long utterance, but we can see that it also contains some 1st person clauses which would go undetected. This illustrates one of the problems arising from the punctuation and style used

5 We acknowledge that the tagger is still at an experimental stage and could benefit both from revision of its set of categories and from performance improvements to gain in informativeness. However, the main aim of its use in the present paper was to assess the outcomes of applying it in its present form to a new dataset as the starting point for further discussion on the usefulness of NLP tools for 19th century data.

by Annie Lough; the tagger is not able to deal with such long multiclausal units, leading to a loss of information when these are encountered in the letters.

(4) and Alice I suppose you are going to school and is at home yet [sic]
(5) I hope you will be at home yet for a long time because it would seem lonesome if you were all gone away
(6) I hope Mother is very well in health
(7) I often thought of the sore hand she had
(8) I hope it will never trouble her again but I wondered why she did not write a few lines any time to me since Easter

All the sentences in (4)-(8) have been correctly identified as 1st person speech acts; the tagger has assigned (5) and (6) to the 1st Person Expression of Feeling category,[6] supported by the use of the phrase *I hope*, and the remainder to the 1st Person Statements category. In sentence (8), we again are faced with the issue of a single sentence containing multiple distinct speech acts, in this case arguably both a statement and an expressive (*I hope it will never…*), but the tagger only assigning a single tag. As well as the potential loss of information, this also raises questions about how we determine the tagger's accuracy – is a 1st Person Statement tag 'wrong' in this case, if it still accurately describes half of the sentence? While this issue does not often arise with workplace emails, which are generally brief and syntactically simple, expanding the tagger to other domains forces us to confront a new scale of 'correctness'.

(9) is Lizzie away in the same place yet

The sentence in (9) is clearly an open question (defined in this study as a non-indirect-request question, cf. Table 4); the SAT, however, has incorrectly marked it as an 'Other' statement. The absence of a question mark – a feature of the original letter – is likely to have contributed to this mistake as it is a strong indicator of an interrogative speech act. Syntax plays a role too, and the parser component of the SAT ought to have recognized that this sentence has the structure of an interrogative, but it failed to identify 'Lizzie' as the subject of the sentence. This also occurs later on in the letter with further interrogatives, shown in (10):

(10) does Nan Deevy live in Meelick yet or did the mill ever start running again
(11) thank the Lord we are living in a peaceful country except Mexico

A different problem arises in categorising sentences such as (11), which, as an exclamation, should be properly tagged as an 'Other' statement. The tagger, instead, assigns it to the '1st person statement' class, influenced of course by the presence of a 1st person clause (*we are living.…*). This example illustrates another challenge in moving from workplace to personal communication. Exclamations are extremely rare in workplace emails, so the SAT has not been exposed to a sufficient number of instances in training to enable it to reliably identify them in the new data.

(12) have this bill exchanged and get yourself a little gift for Christmas from me and if you are not all gone to the Front for war I hope you will write a few lines soon and let me know how you all are

Sentence (12) is the final sentence of the letter. It contains a sequence of directives, which the SAT has correctly categorised, based on features such as the imperative forms of the verbs. As noted above, directives are very rare in these letters. A closer look at the data reveals that the majority of these are instructing the recipient to either write back, send news about particular people or events, or enjoy the material goods sent over by the writer – in other words, actions that maintain or strengthen the bonds between the family members.

The line-by-line discussion of letter 19 provided a good overview of the areas in which the SAT is strongest and the issues that arise in applying it to this dataset. We also carried out an error analysis on the subset of the data identified as incorrectly tagged, to identify any further error trends that can be addressed

6 We recall that this category 'includes any articulation of feelings of personal sentiment such as apologies, joy, congratulations, and so on' (De Felice et al. 2013: 80).

in future versions of the program. In many cases, the cause lies in syntactic features which are typically associated with a specific speech act. For example, (13) and (14) were incorrectly tagged as directives, while (15) was incorrectly tagged as a 1st person commitment. They contain the sequence *you / I + modal verb*, which is often found in those speech act categories (cf. *it would be great if you could do this for me* or *I could finish the report tonight*). The strong association between the syntactic construction and the speech act category leads the SAT to overgeneralise in cases such as these. Constructions absent in the training data also lead to misclassification. This is illustrated by example (16), which has been tagged as a 1st person statement rather than as an expressive. This is due to the use of the construction *feel + adjective*, which does not occur in the training data – expressives in that dataset typically only use the verb *to be + adjective*, so the SAT was unable to recognise this as an expressive.

(13) you might live there sometime yourself yet
(14) don't know when you might get them
(15) they were the newest ones I could get
(16) we all felt real bad about her

Tagger bias can also surface in relation to individual words which in the training data have a particular association with a single speech act. We can see this in (17) and (18), which have been tagged as directive and 1st person expressive respectively. In both cases, a single word (*write* and *sorry*) is the source of the error, as in the training data its presence is a feature of those speech act categories, pushing the SAT towards the incorrect tag.

(17) does Julia write to Maggie at all
(18) he is very well and was very sorry for you

Our brief overview of the main types of tagging errors shows that, while there is no single reason for the tagger's errors in performance, linguistic differences between the training data and the Lough letters are at the source of most of them. As we continue to explore the applicability of this tool to datasets other than workplace emails, we can address this issue by providing new sources of training data to expand the tagger's capabilities, resulting in a domain-aware tool of greater precision.

6. A closer examination of FPFs

In the Annie Lough collection, 284 sentences have been tagged as expressions of feeling (FPFs). As we have seen above, our sample shows that FPFs are identified with 83.5% accuracy, so we believe that this analysis is fairly representative of the category. Table 8 shows the word count as well as the number of FPFs for each letter (the 'Av. FPF' column gives the number of FPFs per 100 words for each letter). The average word count for Annie's letters is 523 with an average of approximately 7 FPFs per letter (or 1.39 FPFs per 100 words).

Table 8. Word counts and FPFs for Annie Lough's letters

Ref.	Word count	FPFs	Av. FPF
1	356	3	0.84
2	480	5	1.04
3	1055	17	1.61
4	487	9	1.85
5	1017	22	**2.16**
6	971	6	0.62
7	208	3	1.44
8	645	5	0.76
9	612	10	1.63
10	541	9	1.66
11	394	6	1.52
12	441	6	1.36

13	365	10	**2.74**
14	431	8	1.86
15	332	6	1.81
16	302	6	1.99
17	632	8	1.27
18	514	11	**2.14**
19	398	3	0.75
20	664	7	1.05
21	884	7	0.79
22	863	9	1.04
23	857	13	1.52
24	469	8	1.71
25	649	6	0.93
26	503	7	1.39
27	396	5	1.26
28	435	4	0.92
29	538	11	**2.04**
30	237	2	0.84
31	206	5	**2.43**
32	870	14	1.61
33	513	7	1.36
34	261	3	1.15
35	400	6	1.50
36	489	5	1.02
37	476	8	1.68
38	207	0	0.00
39	307	4	1.30

Table 8 shows that some letters contain noticeably more FPFs than others. Letters (5), (13), (18), (29) and (31), for instance, contain an average of two FPFs per 100 words (underlined and emboldened in the 'Av. FPF' column). A closer examination reveals that letters (5) and (31) centre on the topic of death and feelings of homesickness and separation. In the remaining letters, Annie writes about, amongst other things, the importance of education (*I hope you keep them to school all you can* (letter (13)), the granting of Home Rule (*I hope with the granting of home rule the coming year that it will be a very prosperous year for dear Ireland and I hope many ones to follow* (letter (18)) and the 'terrible time' in Ireland after World War I (*and I am glad She is so good and smart but so long as you both lived through that terrible time after the war you wont mind it now* (letter (29)).

Examining all FPFs together, it appeared that certain lemmas tend to occur repeatedly: HOPE, GLAD, SORRY, WISH, SURE, THINK, SUPPOSE, KNOW and LOVE.[7] In Table 9, the column 'Freq. FPF' shows how many of the FPFs contain these lemmas. The 'Raw Freq.' column states how many times the lemma occurs across all FPFs. So, for example, the lemma HOPE can be found in 148 out of 284 FPFs and has an overall frequency of 157. Sometimes a lemma occurs just once in an FPF (*Dear Sister I **hope** things go along at the mill as when the old lady was alive*), sometimes it occurs more than once (*I **hope** you and Kate will be very comfortable for the winter and I **hope** you will have your by Christmas*) and sometimes there is a combination of these lemmas (*I **hope** Lizzie is very well and I **wish** she would come home to see you often*). Only 28 out of the 284 FPFs do not contain one or more of the lemmas listed in Table 9.

Table 9. Most frequent lemmas in FPFs in the Annie Lough letters

Lemma	**Freq. FPF**	**Raw Freq.**
HOPE	148	157

7 It should be noted that some of these lemmas (HOPE, GLAD, SORRY, WISH) are used by the tagger to identify expressives in particular, so this relationship is not entirely surprising. We thank one of the anonymous reviewers for encouraging us to explain this finding more clearly.

GLAD	43	44
SORRY	20	21
WISH	18	20
SURE	18	18
THINK	18	19
SUPPOSE	11	11
KNOW	12	12
LOVE	7	7

One thing that stood out when examining the list in Table 9 is that several items (HOPE, WISH, THINK, SUPPOSE, GLAD and SURE) have the potential to realise projection structures. In systemic functional grammar (e.g. Halliday & Matthiessen 2004) projection structures consist of two main components: the *projecting* clause (***I hope***) and the *projected* clause (***you*** *will write*).[8] In these structures the primary (projecting) clause sets up the secondary (projected) clause as the representation of the content of either what is thought, or what is said (Halliday & Matthiessen 2004: 377). Halliday and Matthiessen make a distinction between the projection of propositions and the projection of proposals as follows:

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

Both propositions and proposals have different response-expecting speech functions and an analysis of these structures, therefore, may reveal something about how the author interacts with their intended recipient and the type of response they expect – whether verbal (to provide information), or non-verbal (to carry out an action).

The most frequent verb in the FPFs being examined here is *hope*. In 148 out of 155 occurrences, the verb *hope* is preceded by the first person pronoun *I*. 54 of the 148 occurrences of *I hope* are part of a formulaic greeting (*My dear Sister Mary I hope these few lines will find you all in great health*), or general enquiries about the wellbeing of various family members (*I hope Maggie and her family are all well*),[9] while 25 occurrences relate to the sending or receiving of letters, remittances and other enclosures (*I enclose you a money note for one pound for you and Annie which I hope you will receive in due time* / I *hope you will get it* / *I hope you will receive this letter by Xmas* etc.). The remaining occurrences of *I hope* are perhaps the most interesting in terms of gaining insight into Annie's preoccupations and beliefs and her role within the family. Although less frequent, these occurrences seem to be more personal and reflexive in nature, showing moments of greatest authenticity, directness, expressiveness, and personal identity.

For example, Annie often writes about her desire for Mary and her children to stay close to one another (see examples (19)-(22), below):

> (19) I hope you will be able to go and see Maggie
> (20) I hope you will always have them near you
> (21) I hope Kate and Annie is well and near home and Alice is home with you
> (22) I hope you will be at home yet for a long time because it would seem lonesome if you were all gone away

In example (19), Annie expresses a desire for Mary to visit her sister Maggie, while in examples (20) and (21) she expresses a desire for Mary and her children to live close to one another. In example (22) – a letter to Annie's niece, Katie – Annie 'hopes' that her niece will remain at home with her mother, Mary, for as long as possible. In these examples, through using the projecting clause *I hope*, what is effectively a

8 Other studies have described these structures as clausal epistemic parentheticals (Thompson & Mulac 1991; Huddleston & Pullum 2002), comment clauses (Quirk et al. 1985; Brinton 2008), or meta-discursive phrases (Ädel 2012).

9 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.

command (*be at home for a long time*) is expressed as a statement (*I hope you will be at home for a long time*). Through presenting a command (usually an imperative) as a statement (usually a declarative) – a process which is described by Halliday and Matthiessen (2004) as mood metaphor – the author is able to personalise the command, thereby opening up the possibility for negotiation and interaction. The recipient (in this case Mary or Katie) may choose to follow up on these (albeit indirect) commands, or not.

Related to the topic of family staying geographically close to one another, Annie also expresses a desire for Mary to never experience the trauma and upheaval of migration:

(23) I hope Mary will never have to part hers
(24) I hope you wont have to part any of your children as Mother had to part with us

In example (23) Annie tells her mother that she hopes Mary will not have to part with her children and in example (24) the same sentiment is expressed in a letter to Mary. Annie, then, appears to view migration as forced separation.

The topic of education is also something that Annie appears to feel strongly about. In examples (25) and (26), below, Annie writes to Mary about the importance of keeping the children in school for as long as possible, while in example (27) she indirectly instructs her nieces (Lizzie and her siblings) not to take their education for granted. In these examples, Annie performs the role of aunt, showing an interest in the future lives of her nieces and nephews.

(25) I hope the children are all well and going to school
(26) I hope you keep them to school all you can
(27) I am sure Lizzie is very smart and I hope they will all make good use of their school days

Work is also a common theme. Annie regularly enquires about her brother-in-law's employment, as well as that of her nieces, typically expressing a desire for them to be a) in employment, b) treated well and c) paid good money (see examples (28)-(30), below):

(28) I hope John will be kept in the work
(29) I hope she gets good pay for it
(30) I hope you both have very good places and nice people to work for

A number of clauses can be categorised under the broad heading of 'womanhood'. Annie writes about childbirth and marriage, expressing a desire for her siblings to have children and for her nieces to get married (see examples (31)-(35)). In these examples, we get a sense of women's roles within the notional familial hierarchy, and how the position of women outside Ireland (in the Diaspora) in effect offers a critique of womanhood in Ireland.

(31) I hope you have a nice baby and that it is good and quiet
(32) I hope the next one will be a boy
(33) I hope some time soon to hear of some of you sister getting married
(34) I hope she will have a boy and I hope Lizzie will have a girl
(35) Mother told me Alice was bridesmaid and I hope soon to hear of some of your Sisters becoming a bride

Finally, there are a few FPFs in which we gain a glimpse into Annie's worldview on issues such as World War I (example (36)), and the granting of Home Rule (examples (37) and (38)).

(36) what do you all think of this terrible war that is raging at the present time and when will it ever end. *I hope* no one from our county was foolish enough to enlist but I know many Irishmen has lost their lives there
(37) *I hope* with the granting of home rule the coming year that it will be a very prosperous year for dear Ireland and *I hope* ones to follow Mr Redmond says in his speech no power on earth can defeat the home rule bill it is a very long struggle and *I hope* those that worked so hard for it will live long to enjoy it but think of all the money went from this country to help the cause (Annie to her sister, 8 December n.d.)

(38) *I hope* the times are more settled in Ireland by this time we get the Irish united every week and there is news from every county In Ireland I dont know how the people can stand the british cruelty…this country is doing a whole lot for ireland and we hope that before long congress will recognise the republic of Ireland. John & I belong to the friends of irish freedom a branch was formed here last year Known as Michael Mallon branche he was one of the victims of Easter week what a horrible time that was

7. Conclusion

The speech act tagger offers a useful starting point for access into migrant letters. Although subject to a margin of error, arising from the differences in language between the modern-day workplace emails for which it was developed and the 19th century letters, the tagger has shown to be a valid tool for research in this domain. Speech act annotation opens up novel possibilities in analysing this type of data. We can easily focus on specific speech acts, giving us information about the main communicative functions of this correspondence – what feelings are expressed, what kinds of things are requested of each other, what information is asked for.

In this paper we looked at different types of speech act, focusing in particular on first person expressions of feeling (FPFs). Being able to identify, extract and analyse all FPFs in a letter collection provides us with useful insights into the author's worldview, but also helps us to understand the various functions of migrant letters (how relationships are changed and maintained as well as how roles within the notional familial hierarchy are performed). In Annie's correspondence, for example, almost all FPFs contain a projection structure. As discussed elsewhere (see Moreton 2015), in migrant letters, and correspondence more generally, projection structures often explicitly speak to the recipient of the letter and have the ability to project the author's expectations, desires or beliefs onto the recipient, thus helping to construct what Thompson and Thetala (1995: 103) describe as 'reader-in-the-text'. Projection structures 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.

For this paper we have just looked at letters by one author – 39 letters from a collection of 99. The Lough collection is part of a much larger body of 5,000 letters. The preprocessing work involved in preparing the Lough letters for analysis using the SAT was somewhat time consuming and future research will examine how some of this work might be (semi-)automated. However, through repeating the process we have described here, using letters by authors from a range of socio-historical, economic and cultural backgrounds, a more comprehensive understanding of the functions of migrant letters may begin to emerge, providing a fuller picture of the language of migrant correspondence. Equally too, this further research may show that the functions that have been identified here need to be expanded or refined as other, more representative ones emerge.

Rachele De Felice, University College London
Emma Moreton, Coventry University

REFERENCES

Ädel, Annelie. 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.

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

Brinton, Laurel. J. 2008. *The comment clause in English. Syntactic origins and pragmatic development*. Cambridge: Cambridge University Press.

Baron, Alistair & Paul Rayson. 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.

Bruce, Susannah U. 2006. *The harp and the eagle: Irish-American volunteers and the Union Army, 1861–1865*. New York: New York University Press.

Clark, Stephen & James Curran. 2007. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics* 33(4), 493–552.

Corrigan, Karen P. 1992. I gcuntas De muin Bearla do na leanbhain': eisimirce agus an Ghaeilge sa naou aois deag. In O'Sullivan (ed.), 143–161.

Curran, James & Stephen Clark. 2003. Language independent NER using a maximum entropy tagger. *Proceedings of the CoNLL Conference*.

Curran, James, Stephen Clark & Johan Bos. 2007. Linguistically motivated large-scale NLP with C&C and Boxer. *Proceedings of the ACL 2007 Demonstration Session*.

De Felice, Rachele & Paul Deane. 2012. *Identifying speech acts in e-mails: Toward automated scoring of the TOEIC(R) e-mail task* (ETS Research Report No. RR-12–16). Princeton, NJ: ETS.

De Felice, Rachele, Jeannique Darby, Anthony Fisher & David Peplow. 2013. A classification scheme for annotating speech acts in a business email corpus. *ICAME Journal* 37, 71–105.

De Felice, Rachele & Emma Moreton. 2015. Introducing the Corpus of Business English Correspondence (COBEC): A resource for the lexicon and pragmatics of Business English. Paper presented at the *36th ICAME Conference*, Trier, 27–31 May.

Elliott, Bruce S., David A. Gerber & Suzanne Sinke (eds.). 2006. *Letters across borders : Epistolary practices of international migrants*. London: Palgrave Macmillan.

Emmons, David M. 1990. *The Butte Irish: Class and ethnicity in an American mining town, 1875–1925*. Urbana: University of Illinois Press.

Huddleston, Rodney & Geoffrey Pullum. 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz & David Tugwell. 2004. The sketch engine. Proceedings from EURALEX, 105–116. Lorient, France.

Kilgarriff, Adam & Iztok Kosem. 2012. Corpus Tools for Lexicographers. In Sylviane Granger & Magali Paquot (eds.), *Electronic lexicography*, 31–56. New York, NY: Oxford University Press. www.sketchengine.co.uk (last accessed on 20 November 2018)

Koos, G. 2001. The Irish hedge schoolmaster in the American backcountry. *New Hibernia Review* 5, 9–26.

Miller, Kerby A. 2008. *Ireland and Irish America: Culture, class, and transatlantic migration*. Dublin: Field Day.

Miller, Kerby A. 1985. *Emigrants and exiles: Ireland and the Irish exodus to North America*. New York: Oxford University Press.

Miller, Kerby A., David N. Doyle & Patricia Kelleher. 1995. For love and liberty: Irish women, migration and domesticity in Ireland and America, 1815–1920. In O'Sullivan (ed.), 54–61.

Miller, Kerby A., Arnold Schrier, Bruce D. Boling & David N. Doyle. 2003. *Irish immigrants in the land of Canaan: Letters and memoirs from Colonial and Revolutionary America, 1675–1815*. New York: Oxford University Press.

Minnen, Guido, John Carroll & Darren Pearce. 2001. Applied morphological processing of English. *Natural Language Engineering* 7(3), 207–223.

Moreton, Emma. 2015. 'I hope you will write': The function of projection structures in a corpus of nineteenth century Irish emigrant correspondence. *Journal of Historical Pragmatics* 16(1), 277–303.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 1996. The Corpus of Early English Correspondence. In Terttu Nevalainen & Helena Raumolin-Brunberg (eds.), *Sociolinguistics and language history:*

Studies based on the Corpus of Early English Correspondence, 39–54. Amsterdam: Rodopi.
Noonan, Alan J. M. 2011. ‘Oh those long months without a word from home’, Migrant letters from mining frontiers, 129–135. Cork: The Boolean, University College Cork. http://publish.ucc.ie/boolean/ (last accessed on 20 November 2018)
O’Sullivan, Patrick (ed.). 1995. *The Irish world wide*. Leicester: Leicester University Press.
Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.
Rayson, Paul 2009. *Wmatrix*. Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/ (last accessed 20 November 2018)
Schrier, Arnold 1958. *Ireland and the Irish emigration, 1850–1900*. Minneapolis: University of Minnesota Press.
Thompson, Geoff & Puleng Thetela. 1995. The sound of one hand clapping: The management of interaction in written discoures. *Text and Talk* 15(1), 103–127.
Thompson, Sandra A. & Anthony Mulac. 1991. A quantitative perspective on the grammaticization of epistemic parentheticals in English. In Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to grammaticalization II*, 313–329. Amsterdam: John Benjamins.