Corpora, Catalogues and Correspondence: The Item-Level Identification and Digitisation of Business Letters for the British Telecom Correspondence Corpus

Morton, R. and Nesi, H.

Corpora, Catalogues and Correspondence: The Item-Level Identification and Digitisation of Business Letters for the British Telecom Correspondence Corpus

by Ralph Morton and Hilary Nesi

Citation

Morton, Ralph and Hilary Nesi. 'Corpora, Catalogues and Correspondence: The Item-Level Identification and Digitisation of Business Letters for the British Telecom Correspondence Corpus'. In: Clare Mills, Michael Pidd and Jessica Williams. *Proceedings of the Digital Humanities Congress 2014*. Studies in the Digital Humanities. Sheffield: HRI Online Publications, 2014.Available online at: <http://www.hrionline.ac.uk/openbook/chapter/dhc2014-morton>

Abstract

This paper explores some of the challenges in working with archive material to produce language corpora. It takes as a case study the *British Telecom Correspondence Corpus* (BTCC) which contains a selection of the letters held in the BT Archives, housed in *Holborn* Telephone Exchange. One of the essential differences between a corpus and an archive is that a corpus is intended to be representative of a language variety. Material makes its way into historical archives in a variety of ways, and whilst they may preserve a breadth of material; archives are not generally collected to be representative, nor are they primarily designed to facilitate linguistic investigation.

Work on the BTCC began as part of a Jisc-funded project to digitise the BT Archives and create a ‘research resource for the higher education sector’ (Hay, 2014:12). The BT Digital Archives became available to the public in July 2013. Our experiences using this resource inform the second half of the

paper, in particular regarding the identification of corpus material and the difficulty in identifying letters at an item level. This leads to a wider discussion of how best to digitise physical archives.

Corpora, Catalogues and Correspondence: The Item-Level Identification and Digitisation of Business Letters for the British Telecom Correspondence Corpus

by Ralph Morton and Hilary Nesi

1. Project Background (British Telecom and *New Connections*)

British Telecom (BT) is the world's oldest communications company, tracing its history back to the formation of the Electric Telegraph Company in 1846. The BT Archives were established in 1986 and contain a vast array of material from the founding of the Electric Telegraph Company through to the present day. Before its privatisation in 1984, British Telecom had been a government department, a nationalised industry and a public corporation. The pre-privatisation material in the archive is public record and BT is obliged to promote access to it. However despite this legal obligation, the physical availability of the texts is restricted in practical terms. Access to the Archive offices in Holborn, London has been, and remains, limited to two days a week. At the time of the project reported in this paper our contacts at the BT Archive were David Hay, Head of Heritage and Archives, and Siân Wynn-Jones, Heritage Collections Manager.

The National Archives catalogue, which uses a similar structure to the BT Archive, is organised at the following levels:

1. **Department** – "the government department, agency or body that created the records".

2. **Series** – "a main grouping of records with a common origin and function or subject matter".

3. **Piece** – "not a single piece of paper; it may be a box, volume, file, roll and so on" .

4. **Item** – "a part of a piece. It can be a bundle, a single document, a file, a sub-file, a pouch, a range of folios and so on" (for more detail see http://nationalarchives.gov.uk/records/citing-documents.htm).

The only difference between this and the structure used by BT is that at the top level, rather than department, the BT material is organised by "broad groups (called 'fonds') that reflect the main chronological periods of the organisation (David Hay, personal correspondence, 10th April 2014). In organising the records in this top-down manner, the aim is to "reflect the nature and structure of the organisation that created them" (IBID).

The *New Connections* project was set up to address the limited accessibility of the archives, by cataloguing, digitising and developing a searchable online archive of almost half a million photographs, images, documents and correspondence assembled by BT over 165 years. The project was a Jisc Company (formerly *Joint Information Systems Committee*) funded collaboration between Coventry University, The National Archives and BT Heritage. Broadly speaking the roles of the partner institutions were that BT provided the material, The National Archives digitised the material, and Coventry University developed the archive website and undertook research projects using the digitised material. The research project undertaken by the English and Languages Department at Coventry University was the construction and analysis of the British Telecom Correspondence Corpus (*BTCC*). The *New Connections* project ran from November 2011 until July 2013, when the Digital Archive was launched. The BTCC was completed in April 2015.

2. Introduction to the British Telecom Correspondence Corpus (BTCC)

The era covered by the BT Archive makes it a potentially fascinating source of business correspondence data. The mid-nineteenth century saw a huge increase in letter writing, facilitated by the introduction of the Penny Post in 1840 (Dossena and Ostade, 2008: 7-8). There was also specifically an increase in business correspondence brought about by the new commercial climate following the Industrial Revolution (Beal, 2004: 116, Del Lungo Camiciotti, 2006:153). Lyda Fens-De Zeeuw's comparison of eighteenth and nineteenth century letter writing manuals finds that the business letter's instruction component, which was non-existent in the eighteenth century

manuals, became ‘prominent’ in the nineteenth century manuals (2008: 189).

Despite its potential as a research resource, there is relatively little business correspondence data from the nineteenth and twentieth centuries in available corpora. Marina Dossena’s Corpus of Nineteenth-century Scottish Correspondence (2004)contains a business element, but as the name suggests, only contains nineteenth century data and focuses on just one variety of English (Scottish). Researchers who have attempted to study the development of correspondence from the nineteenth to the twentieth century have often relied on model letters from letter writing manuals, as did Del Lungo Camiciotti (2006), or have been unable to make quantitative comparisons due to a lack of twentieth century data, as was the case with Kytö and Smitterberg’s (2006) study.

The BT Archives are well suited to corpus study as ‘natural data not produced for linguistic analysis’ (Stubbs, 2007: 130). The texts are a product of a very specific context and so may not represent the language of correspondence across all types of business settings. Nevertheless the corpus has great value as a means of tracking business English’s development from the nineteenth century through to the late twentieth century.

3. Selection of Material for the BT Digital Archive

BT’s initial selection of material for the Digital Archive involved consultation with an external Advisory Group, but was ultimately based on the BT Archive team’s “knowledge of existing and potential research value in the collection” (Hay, 2014:8). It was decided that whole files should be scanned rather than specific documents, so that “higher levels of credibility and lower levels of personal subjectivity would be achieved” (IBID). The chosen files, considered to have the most ‘research value’, contained material representing:

- Technology milestones
- National events
- International reach
- Government policy
- Industrial relations
- Diversity
- Iconic products & services

By the time we started constructing the corpus at Coventry University, the digitisation process was already underway: the files had been selected and the National Archives had started preparing and scanning them.

4. British Telecom Correspondence Corpus

At the time of the last audit of the BT Archives in 2009, there were 1,761 meters of folders containing 'registry/subject files' (the category of file that correspondence can be found in) (Sian Wynn Jones, personal communication, 5th March 2013). It is not currently known how many letters are contained within the archives. Our proposal to Jisc as part of the *New Connections* project was to construct a corpus containing around 500 letters, selected from all the documents scanned for inclusion in the Digital Archives. It was felt that this was a realistic number given the limited time and resources. Weplanned to include a spread of letters from across the fifteen decades of material represented in the archive. With this in mind, BT agreed to provide Coventry University with scans of the contents of around 1,000 subject/registry folders. These folders were provided solely on the basis of being sufficient for the linguistics part of the project i.e. it was thought that 1,000 folders would contain at least 500 letters from various points across BT's history.

Thirteen thousand scans of individual pages of these subject/registry files were delivered to Coventry University on a hard-drive in May 2012, named and organised according to the BT Archive folder/file 'finding numbers'. At this point the only available metadata for the documents (folder contents, participants, year, topic etc.) was recorded on the first scanned page of each folder (or 'file', using archive terminology). The amount of information recorded in these descriptions varied greatly from file to file. Some had detailed contents lists while others had descriptions such as 'miscellaneous papers'. Due to the patchy nature of the existing metadata there was no easy way to identify which files contained letters, and if so how many; the only feasible way to extract the documents we wanted was to manually examine all 13,000 scans.

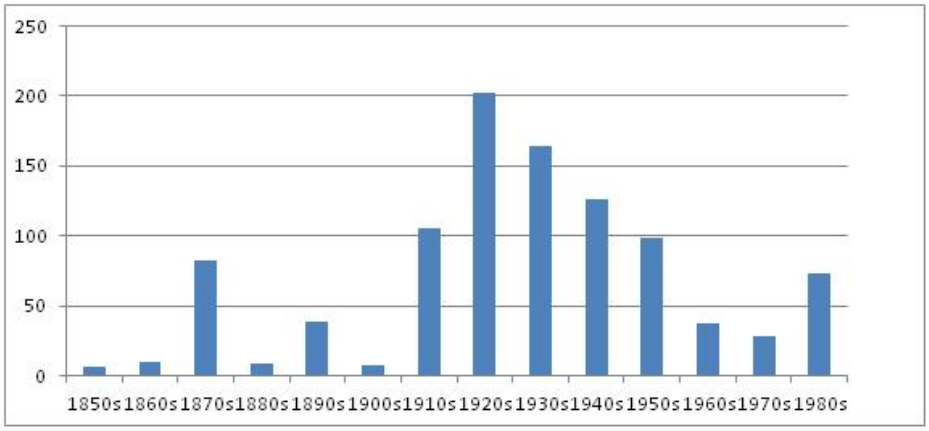
5. Defining and Identifying 'letters'

The criteria employed for identifying ‘letters’ were based on the typical formal features of a letter as recommended in writing guides for official correspondence (e.g. Nesfield, 1917, Thomson 1972). We decided that for our purposes a ‘letter’ should contain:

1. an address (full or partial)
2. a date
3. an opening salutation
4. a closing salutation
5. a signature

Additionally, only finished letters were selected, as opposed to drafts, memos or other text-types that are also categorised by BT as ‘subject/registry’ files.

Through manual examination of the 13,000 scans, 991 ‘letters’ were identified as meeting the formal conditions identified above. These letters were then grouped by year and decade. The earliest letter identified was from 1853 and the latest was from 1982, so the corpus was able to span 129 of the 137 years of public records contained in the archives, at least to some degree. However, these 991 letters were very unevenly spread across the decades.



6. Corpus vs. Archive

Ideally to create an ‘internally contrastive’ (Sinclair, 2005: 3.1) historical corpus, an even distribution of data across the period is preferred. A perfectly balanced corpus of 500 letters would contain 36 letters per decade. However as William Labov noted “historical documents survive by chance, not by design, and the selection that is available is the product of an unpredictable series of historical accidents” (2001:11). Even given the relatively recent nature of the historical data in the BT archive, some decades were far better represented than others in terms of surviving correspondence. In order to achieve a more even distribution, the decision was made to include in the 500-letter corpus every letter from under-represented decades, and a more-or-less even selection from each of the remaining decades.

This selection process gave us more control over the sample of material included in the corpus and the type of corpus we wanted to create. Leech (1991:11) argued that ultimately, the difference between an archive and a corpus must be that the latter is designed or required for a particular 'representative' function’ (see also Sinclair, 2005). However as Hunston (2002: 28) points out, “the problem is that ‘being representative’ inevitably involves knowing what the character of the whole is” (see also Biber, 1993: 243).

As we have seen, the ‘character of the whole’ of the BT Archives is well

defined in terms important developments in the company's history. Folder contents are grouped and catalogued in terms of "the context and the function of the folder in relation to the history of BT and its predecessors" (Sian Wynn Jones, personal communication, 5th March 2013). However this focus on Fond and Series level organisation means documents are not typically well described at an item level. This poses a challenge in the selection of linguistic data, as ideally we would want to have some idea of how many letters the archive contains, and the population of authors and recipients involved.

Given the quantity of folders containing 'registry/subject files' (a little over a mile in shelf space) it is not currently feasible to explore how many letters are contained within the BT archive.

7. Sampling Material for the BTCC

The way we addressed this lack of definition of the population of authors in the BT Archives was to sample in a way normally reserved for spoken language which "exists in unknowable quantities and in an unknowable range of varieties" (Hunston, 2002: 29). We selected a number of categories that we wanted to control and sampled according to them: a purposive approach.

For the BTCC, the primary concern was to include roughly equal number of letters per decade to make it possible to study linguistic change over this period of business correspondence. We also wanted to include as many authors as possible. It was felt that the variety of authors in the archive material is one of its main strengths as a source of linguistic data, as it nullifies to some degree the effect of individual authors' stylistic idiosyncrasies. An attempt was made to include both historically prominent figures and ordinary employees as authors and to include some chains of correspondence to allow for the study of traceable interactive elements (Dossena, 2004: 198). As far as possible we also wanted to represent both male and female authors, although there were very few female authors in the sample of texts we were given.

Both handwritten and typewritten letters were included in the corpus; handwritten letters predominated in the first four decades (1850-1890) but some letters continued to be handwritten even up to the 1980s. We included

samples of handwritten letters from later decades as handwriting sometimes indicated that the author was writing in a non-official capacity. One hundred and fifty of the typewritten letters were successfully transformed to plain text at The National Archives, using Optical Character Recognition software. Carbon-copy typewritten documents and those that had not been well conserved had to be manually transcribed, as did all the handwritten letters.

Part of the stated purpose of the *New Connections* project was to help improve the cataloguing of the BT collections and although archive-wide metadata is still lacking, the BTCC has begun to give us an idea of the sorts of correspondence that the archive contains. Correspondence has the advantage of being particularly rich with metadata; almost all the letters in the BTCC provide names, dates and geographical locations in the form of postal addresses, making it possible for us to piece together contexts for both the texts and their authors. In this sense the corpus construction is exploratory and demonstrates the mutually beneficial nature of projects like this, where a research institution is given access to archive resourc0es, and in turn can improve the archive's understanding of its own content.

The massive scale of the archive means that the corpus could potentially be supplemented with additional material at a later date, although the problems outlined above regarding letter identification and extraction would continue to apply.

8. Detailed Metadata Collection

Due to time constraints, the metadata used for the initial selection of the letters was fairly basic. More detailed metadata has since been collected through a close examination of the letters. The criteria that qualified a text as a letter mean that at very least an address, a date and a signature were available. Some opening salutations also contained recipient information. The selection criterion 'signature' did not always mean that we were able to extract a name, as some were illegible. However only 10 of 379 names were completely illegible, and it was still possible to give these authors identifiers, and extract other relevant information about them from the letters.

In addition to names and locations, the job titles of authors and recipients and the names of the companies that they worked for were recorded. In terms of the linguistic analysis, this provides crucial information about the

status of and relationship between correspondents. On a more general level it also helps us to build up a picture of the wider business network of which BT was a part. Author gender was also recorded, as was the format of the letter (typed, handwritten, typed copy, handwritten copy, form letter, handwritten telemessage, typed telemessage, typed w/handwritten note).

Text Encoding Initiative (TEI)–compliant XML was used for corpus annotation. This allows for the annotation of bibliographic and contextual information in the document 'header' along with textual and structural annotation in the body of the text. All of the metadata detailed above was recorded in the header along with: a description of the project, information about the file repository (BT Archives), original file references, transcription practice, format, pragmatic classification and availability information.

Correspondence-specific contextual information was annotated using the newly developed *Text Encoding Initiative* (TEI) Correspondence Guidelines (https://github.com/TEI-Correspondence-SIG/correspDesc).

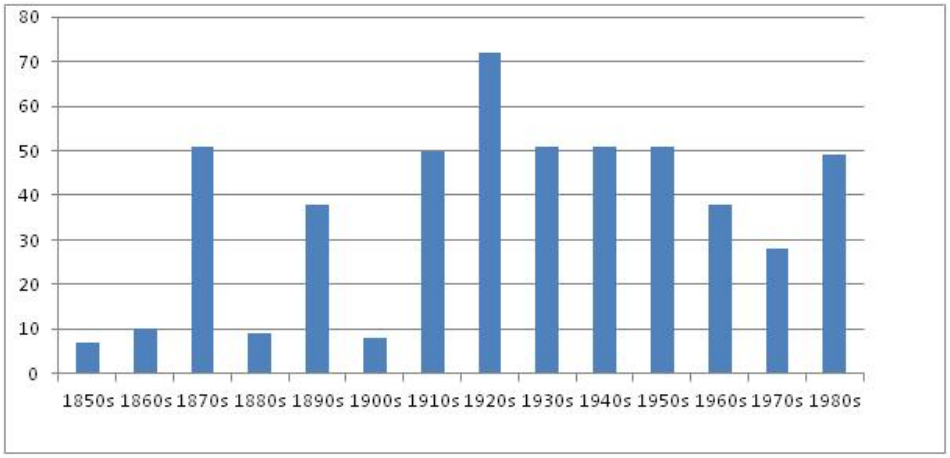
Below is an example of the structure of the correspondence-specific header information for one of the letters in the *BTCC*

```
<profileDesc>
   <correspDesc>
      < correspAction type="sending">
       <persName ref="HSW">
         <forename>Henry</forename>
         <surname>Schütz-Wilson</surname>
         <roleName>Assistant Secretary</roleName>
     </persName>
     <placeName>
            <orgName type="Company">The Electric &
International Telegraph
             Company</orgName>
         <address>
             <street>Telegraph Street</street>
             <settlement>London</settlement>
             <postCode>EC</postCode>
         </address>
     </placeName>
   <correspAction>
```

```
  <correspAction type="receiving">
       <persName ref=”EB”>

          <forename>Edward</forename>

          <forename>Brailsford</forename>
           <surname>Bright</surname>
           <roleName>Secretary</roleName>

      </persName>
       <placeName>

             <orgName type="Company">British and Irish
Magnetic Telegraph Co.</orgName>

      </placeName>
       <date>N/A</date>
     </correspAction>
   </correspDesc>
 </profileDesc>
```

9. Rejoining the process as an archive user: the launch of the BT Digital Archive

Figure 2: Number of letters per decade – 512 letters.

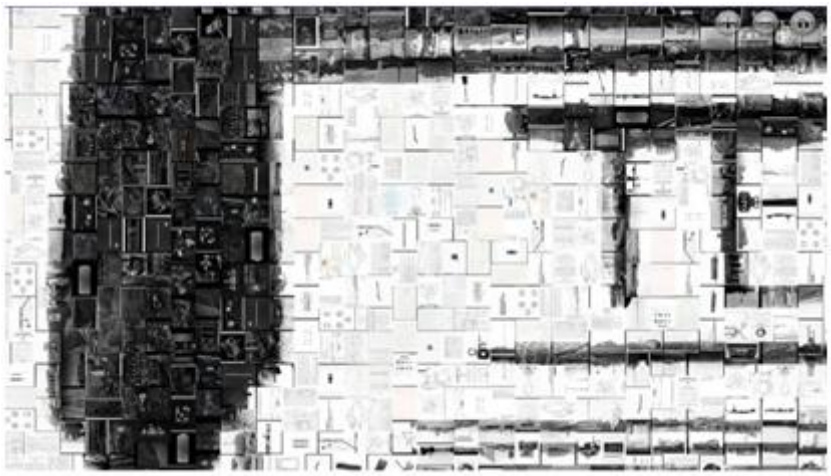


The corpus was at this stage relatively balanced, although there was still scope for improvement in terms of distribution across the decades, as can be seen in Figure 2.

The BT Digital Archives website (http://www.digitalarchives.bt.com/web/arena) went live in July 2013 and contained all of the material scanned by the National Archives. This presented the possibility of searching the archives for additional letters to balance the remaining under-represented decades.

10. “Serendipitous searching”

On entry to the website users are presented with a mosaic made up of thousands of photographs and scanned documents which come together to form an image from the archive (see Figure 3). As you click on the mosaic you zoom in to the image (see Figure 4) until you reach a single record page accompanied by (where available) a title, finding number and a date. As such it allows for “serendipitous searching” (Hay, 2014:13) of the digital archive.



Targeted searches are more problematic, however. Documents are not categorised by type so searches within record pages for ‘letters’ and ‘correspondence’ return results from file and series names and descriptions, rather than individual items. Furthermore there is considerable variation in the amount of detail these descriptions contain. For example a search for record items using the search term “correspondence” directs users to the folder '*Correspondence on Marconi's experiments into wireless telegraphy and the establishment of the Marconi Company*'.

The folder contains 69 pages of information, some but not all of which is correspondence. The only way to find individual scans of pages of letters is by scrolling through each page of the folder. While it is not such an arduous task to sift through this one folder, a search for ‘letter’ returns 273 pages of results of folders the names or descriptions of which contain the word ‘letter’, and a search for ‘correspondence’ returns 508 pages of search results.

Searching the records by date can also be difficult. For example, the *Correspondence on Marconi’s experiments…* folder cited above is dated 1896-1909. To create a more event spread of letters across the corpus we required additional letters from the 1910s but not from the 1890s, so the results were both too limited and insufficiently detailed

However, although finding individual documents remains an imprecise process, access to material through the BT Digital Archives is undeniably an improvement, allowing users to see documents that could only previously have been viewed at the BT Archives in Holborn.

11. Using Finding Numbers to Describe the Corpus

One form of item-level document search made possible through the BT Digital Archives is search by finding number. Finding numbers contain information regarding the fond, series and file that an item belongs to.

To take the example, *TCB_273_2_27,* from the *BTCC* data, “TCB” locates the record as part of the Fond *Post Office Telegraph and Telephone Service, 1868-1969*, within this ‘273’ is the series *Correspondence regarding wireless telegraphy,* and within this seriesfiles ‘2’ and ‘3’ relate to *Correspondence on Marconi's experiments into wireless telegraphy and the establishment of the Marconi Company.*

As we had already received 13,000 items of material from BT, each named according to its finding number, we were able to use the file names to search the BT Digital Archives and identify the topic of each letter, as defined by BT. This was a great help in describing the nature of the documents in the BTCC. The letters relate to a wide variety of topics and we would not have had the time to develop our own taxonomy. The following are the fifteen most frequent topics in the corpus (the number of letters on each topic is indicated in brackets):

Guglielmo Marconi's relationship with William Preece and experiments with wireless telegraphy (40),

N/A (33),

Correspondence with government – miscellaneous (28),

Correspondence on Marconi's experiments into wireless telegraphy and the establishment of the Marconi Company (27),

Transatlantic Wireless Telegraphy (25),

Rental of Professor Graham Bell's instruments by the Post Office (25),

Network competition (24),

Trials of Thomas Edison's modification of George Little's automatic telegraph system (23),

Arrangements for Post Office staff during the Second World War (22),

Cooperation with NASA (20),

Telephones: 'Tim' speaking clock Miss Cain selected, 100 Guinea fee (17),

Telegraph companies' records discovered in vault 61 of the General Post Office West (17),

Employment of coloured people (17),

Public telephone kiosks: competition for improved design, Sir G Scott's model adopted (17),

Telegrams (15),

Cost of Telegraph Service: Analysis of Telegraph Commercial Accounts 1925 – 1937 (14),

While not all of these are ideal as topic names, many of the finding number searches were useful for our purposes and helped relate the material to the rest of the archive. However, most users of the Digital Archive will not have access to finding numbers when conducting searches, and so face the cyclical problem that if you do not already have the finding number for a document it is very difficult to find it. There were also 33 letters for which we had the finding number but were not able to find information at the fond, series, file or item level in the Digital Archives.

12. Search Results from the BT Digital Archive

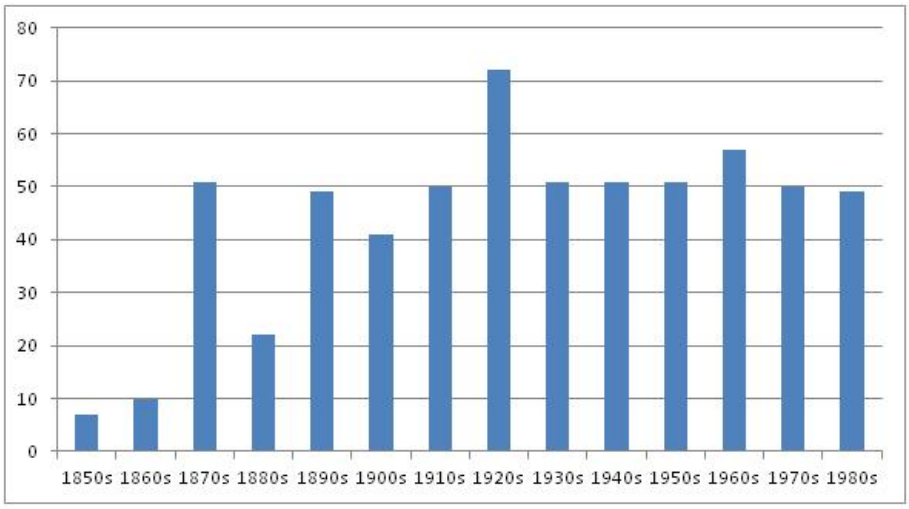
Despite the imprecise nature of the search process, it did point us towards a number of folders that contained letters. A long series of queries enabled us to identify the following quantities of letters in the ten most recent decades in the Archive:

1870s – 1, 1880s – 14, 1890s – 11, 1900s – 34, *1910s – 35, 1920s – 13, 1940s – 4*, 1960s – 20, 1970s – 30, *1980s – 1.*

Some of these decades (indicated in italics) were already sufficiently well represented in the BTCC. However it was possible to resolve some of the imbalances across decades. After the addition of the material from the BT Digital Archive the total number of letters increased to 611. These letters'

distribution across the fourteen decades can be seen in Figure 5.

Figure 5: Number of letters per decade – 611 letters.



13. To Digitise More or Digitise Differently?

Questions about the extent of digitisation are likely to be an ongoing concern for large physical archives. Chris Barnes, from the Advice and Records Knowledge Department of The National Archives, advised that:

> *"Most record series are not easily identifiable on our catalogue due to their chronological arrangement or lack of description at item level. Some records, such as those of the treasury are impenetrable from outside TNA as you must consult a series of indexes only available in the reading room itself"*(personal communication, March 2014)

There is a danger that if digital archives merely replicate the physical archive they will inherit much of the impenetrability of the physical archives, without the indexes or archive employees to provide the same level of assistance. Born-digital resources are typically very easy to access at the item level. The difficulty of identifying relevant materials within the traditional archive structure may be a significant barrier to the wider use of

archive material for academic research, and this would be a shame because archives like that of BT and the record house at the National Archives are potentially very interesting sources of linguistic data.

A key theme that emerged from the *Digital Humanities Congress 2014* was the question of whether to digitise more or to digitise differently. An impressive variety of digital resources were presented at the conference, but uncertainty was expressed by some of those creating these resources as to what exactly researchers hoped to use the resources for. A better understanding of the intended digital archive user would be helpful in focusing the creation of digital resources, but in focusing on one set of priorities others are inevitably ignored. Each project team has choices to make in terms of what should be digitised.

The Post Office, of which BT used to be a part, chose to focus on the digitisation of their archive catalogue. Around eighty four per cent of their archive has been catalogued (Gavin McGuffie, Archive Catalogue and Project Manager, personal communication, 5th August, 2014) and this catalogue is available to search online (http://www.postalheritage.org.uk/). It is also possible to filter results by subject, creator, format, location and level within the catalogue. However in the Post Office digital archive only around fifteen per cent of records have an image, and the bulk of these digitised images are currently photographs and sheets of stamps. In practical terms then, despite its limited searchability, the BT Digital Archive has made more images of documents available than the Post Office Heritage website.

Each digitisation project is likely to have its own aims and there may not be a single approach that suits all; nevertheless the development of standard guidelines for digitising particular document types could help inform project managers' digitising decisions.

14. Digitising Letter Collections: What Information do you Need?

Recently the AHRC-funded networking scheme *Digitising Experiences of Migration* set out to develop a system of mark-up for emigrant letter collections. The project, discussed at the Digital Humanities Congress 2014, brought together scholars from different disciplines with the aim of "initiating the process of interconnecting resources to encourage cross-

disciplinary research". The workshops held in Utrecht, Lancaster and Omagh were set up "to understand and map the linguistic, structural, discoursal, contextual and physical properties of the letters that each stakeholder group is working with".

In terms of contextual information the fundamental properties of letter collections were found to be author and recipient information, location and date. Whatever other complex systems individual researchers and disciplines employed in their study of correspondence, standardised information relating to these areas was agreed upon as a basis for the greater interconnectivity of digital resources. This discussion fed into the development of the *Text Encoding Initiative's* Correspondence Special Interest Group's proposed guidelines.

With compatible metadata it is possible to create resources such as *correspSearch* http://correspsearch.bbaw.de/ which is based on the TEI *correspDesc* structure and brings together the metadata of four scholarly editions of German letter collections, making it possible to search through these collections by sender, addressee, place and date. Extracting this sort of basic item level information when digitising archived letters could not only vastly improve the searchability of individual items but also creates exciting opportunities in terms of linking letter connections and providing a clearer impression of the personal and business networks that the correspondence helped to maintain.

15. Enhancing Item Level Archive Description Through Collaboration

Archive-wide production of metadata on an individual item level would be an overwhelming task for any one institution or research project. However one of the benefits of making archive material available for research is that researchers will inevitably improve the description/metadata for individual files when investigating documents. As the National Archives advise on their website

"Traditionally, citation of our records was done at piece level (generally the unit of production at Kew). As itemisation has become common, in order to enhance the descriptions of our records and to enable digitisation, researchers may wish to cite an individual item within a piece."

http://nationalarchives.gov.uk/records/citing-documents.htm

Digital Archives also offer the possibility of crowd-sourced tagging whereby individuals interested in a digital archive can ‘tag’ an individual page or document, making it easier for subsequent researchers to locate relevant records. This feature is currently available to registered users of the BT Digital Archives.

Doherty (2007) supported the creation and use of authority records, particularly in a business context where they can be used to describe the changing corporate identity of companies like BT. In producing the files for the British Telecom Correspondence Corpus we have extracted information that could form the basis of authority files. We have also preserved some of the document history from the BT Digital Archive in the form of the old POST class references.

The new file references for the *BTCC* encapsulate basic information about the date, author and recipient of each letter. It is hoped that these file names will be less impenetrable than those used in the original archive, and that in combination with the TEI-encoded metadata they will provide sufficient information for any future linguist or archive researcher in relation to the questions ‘Who?’ ‘What?’ ‘When?’, and ‘Where’?

16. Conclusion

From a linguistic perspective the digitisation of archives containing authentic language data is very valuable. Working with archive material to create linguistic resources can prove problematic as archives are organised and catalogued in relation to historical periods and events rather than at item level. Archives are not collected to be representative, nor are they primarily designed to facilitate linguistic investigation; they are arranged according to the requirements of the original archive creators and users.

This article suggests that the digitisation of archives presents opportunities to engage with material in new ways. To be widely used by other interested parties they should also be compatible with other, perhaps previously unforeseen requirements. It may be that document type should be taken into consideration in digitisation to maximize the usefulness of that document in a digital form.

Whatever the present limitations of the BT Digital Archives, even working with what is presumably a tiny sample of the correspondence contained in the archive overall, we start to get an idea what sort of material the archive contains. This information feeds back into the understanding of the history of BT and its predecessors, and 'file' and 'series' descriptions could potentially be supplemented by item level description through collecting and encoding contextual information.

We are currently in discussions with the Post Office to expand the corpus to include material from their archive. This will strengthen the corpus and bring documents with a shared institutional history back together to shed light on their collections' linguistic as well as historical past.

References

Beal, J.C. (2004) *English in modern times.* London, Arnold.

Biber, D. (1993) 'Representativeness in Corpus Design', Literary and Linguistic Computing, 8(4), 243-257.

Del Lungo Camiciotti, G. (2006a) 'Conduct yourself towards all persons on every occasion with civility and in a wise and prudent manner; this will render you esteemed: Stance features in nineteenth century business letters' in Dossena, M. and Fitzmaurice, S. (eds.) *Business and Official Correspondence: Historical Investigations,* Bern, Peter Lang, pp 153-174.

Doherty, T. (2007) 'Who, what, when, why?: ISAAR (CPF), the forgotten standard?' in *Business Archives: Principles and Practice*, 87.

Dossena, M. (2004) 'Towards a corpus of nineteenth-century Scottish correspondence' *Linguistica e Filologia* 18, 195-214.

Dossena M. And Tieken-Boon van Ostade, I. (eds.) (2008) *Studies in Late Modern English Correspondence: Methodology and Data,* Bern. Peter Lang.

Fens-de Zeeuw, L. (2008) The Letter-Writing Manual in the Eighteenth and Nineteenth Centuries: From Polite to Practical. In Dossena, M., Tieken-Boon van Ostade, I (eds.) *Studies in Late Modern English Correspondence:*

Methodology and Data. Bern, Peter Lang, pp 163-192.

Hay, D. (2014), 'New Connections: The BT Digital Archives Project' paper presented at the *Archives and Cultural Industries Conference*, Girona, 11th-15th October.

Hunston, S. (2002) *Corpora in Applied Linguistics* Cambridge, Cambridge University Press.

Kytö M. & Smitterberg S. (2006) '19th-Century English: An Age of Stability or a Period of Change' in *Corpus Based Studies of Diachronic English* ed. by Facchinetti R. & Rissanen M. Bern, Peter Lang, pp 199-230.

Labov, W. (2001) *Principles of Linguistic Change: Internal Factors* Oxford, Blackwell Publishing.

Leech, G. (1991) 'The state of the art in corpus linguistics' In Aijmer, K. & Altenberg, B. (eds.) *English Corpus Linguistics: Studies in honour of Jan Svartvik,* 8-29.

Nesfield, J. (1917) *Junior Course of English Composition,* London, MacMillan and Co. Limited.

Sinclair, J. (2005) 'Corpus and Text: Basic Principles' in *Developing Linguistic Corpora: A Guide to Good Practice* ed. Martin Wynne 1-16 http://ahds.ac.uk/linguistic-corpora/.

Tantony, M.,(2013) http://postalheritage.wordpress.com/2013/09/25/cataloguing-archives-in-four-very-easy-steps/.

Thomson. K.G. (1972) *The Pan Book of Letter Writing* (9th edition), London, Pan.