Public Archives as a Source of Historical Linguistic Data: The Construction and Analysis of the *British Telecom Correspondence Corpus*

Ralph Morton (Coventry University, UK)

Project Background

The BT Archives house the records of British Telecom, the world's oldest telecommunications company, which traces its history back to the formation of the Electric Telegraphy Company in 1846. Prior to its privatisation in 1984, BT was a public corporation (and before that a government department) and as a result all of the pre-privatisation material in the archives is in the public domain, making it ideal for academic research. Despite this legal availability, however, the physical availability of material in the archive was limited to two days a week in an archive space in Holborn, London. In 2011 the 'New Connections' project was set up with the aim of making around half a million items from the public archives of British Telecom available in a new digital archive. As part of 'New Connections', three academic research projects were funded, one of which was the creation and analysis of the British Telecom Correspondence Corpus (BTCC).

The era that the archive covers makes it a potentially fascinating source of data for the linguistic study of business correspondence. The mid-nineteenth to late-twentieth century is a crucial period in the development of English business correspondence as the amount of business being conducted by letter increased massively during this period as a result of the Industrial Revolution, the introduction of the Penny Post, and increased access to education both in schools and through composition grammar guides. Despite its importance in the development of business correspondence, this period has received relatively little attention. The aim of constructing the British Telecom Correspondence Corpus was to start addressing this gap in available linguistic data and enable studies into the development of business correspondence from the mid-nineteenth to late-twentieth century.

Creating the Corpus

The first major challenge in creating the corpus was to identify letters. The material for the digital archive had been preselected by British Telecom, and was digitised and delivered as image scans to Coventry University by The National Archives. In this sense, the initial data collection had already been done. However, the digitised files were labelled and organised according to British Telecom and Post Office archive finding numbers (e.g. TCB 473/P 10045) and contained little or no metadata. The only way to identify letters was, firstly, to set criteria as to what constituted a 'letter' for our purposes, then search through around 13,000 individual scans manually. This first phase of searching identified just over 500 letters. Once the digital BT Digital Archive launched it was possible to search for additional letters but again the lack of item-level metadata meant that similar challenges persisted. Overall 612 letters were selected for the corpus, which in its current state contains just over 130,000 words.

The letters are distributed relatively evenly across the fourteen decades represented. An attempt was made to include a wide a variety of authors, occupations, and companies as possible to get as wide a representation of the available material as possible. The correspondence is mostly written in British English, though as British Telecom and the Post Office did a great deal of international business there are letters from many different locations, perhaps most notably America from which there are letters regarding, for example, the first trans-Atlantic Telephone calls and co-operation over satellite testing. The authors are largely male despite an initial hope that the corpus might be more balanced in terms of gender. This imbalance seems largely due to the roles that women typically held in telecommunications companies during this period which did not generally involve authoring correspondence. The data was transcribed through a mixture of manual transcription and OCR (Optical Character Recognition) scans of type-written material.

Finally, the letters were classified in relation to the overall pragmatic function they serve. A relatively small list of ten functions (such as *Applications*, *Requests, Commissives, Offers*…) was used to try and limit the effects of data scarcity, while offering a starting point for comparison of the various sorts of letters represented in the corpus. The classification also offered an additional way of interpreting the purely quantitative results (for example the n-gram '*let me know*' appears most frequently in *Queries*).

Methods of Analysis

The analysis of the corpus was primarily data-driven, that is to as few preconceptions as possible were made about the data, and I worked from the starting point that, as Stubbs put it, 'repeated events are significant' (2007: 130). This approach seemed consistent with the core principles of corpus linguistics, and appropriate for an essentially exploratory study. I generated a list of n-grams of between three and six words that appeared more than 25 times in the corpus as a whole, and examined each for stability and change with regard to frequency and function over the timeline of the corpus. In addition to this keywords were generated using each decade as a sub-corpus and comparing against the whole corpus. These keywords were organised according to the categories outlined by Scott (2012): proper names, indicators of 'aboutness', and potential indicators of style.

The majority of the keywords (66%) were indicators of 'aboutness'. Many of these results were interesting in terms of topic but of limited linguistic interest and so were used to help contextualise findings in terms of topic. Each of the potential style-marker keywords, which made up 12% of the keywords overall, were examined in context for how they were used and whether the patterns in which they appeared changed over time.

Some Initial Findings: Corporate Identity

<u>Decline in Deference: Formalised distance to formalised friendliness</u>

One clear pattern highlighted by the quantitative results is a decline in overtly polite and deferential terms of address. Seven of the most frequent n-grams are variations

on the formulaic closing '*your obedient servant'*, which is most often preceded by another frequent three-word n-gram *'I am sir',* the seventh most frequent three-to-six-word n-gram in the BTCC. Overall these formulas decline in frequency in the early-mid twentieth century and disappear from the corpus in the late 1950s.

We also see a corresponding decline in the use of formal opening terms of address such as *'Sir'* in favour of the use of named recipients (e.g. '*Dear Ken*'). In the latter half of the twentieth century '*Dear [first name]*' is increasingly the most popular opening formula, and '*yours sincerely*' dominates closing formulas. Similarly, where authors in the late nineteenth and early twentieth century tended to manage the exchange of letters with phrases such as '*with reference to your letter*', increasingly authors write '*thank you for your letter*'. This seems to be part of a wider decline in more negatively polite (Brown and Levinson, 1987) forms in favour of positively polite forms of address, which historically have been more typical of personal correspondence. While these forms are familiar on the surface, however, they also occur in increasingly standardised forms with no variation to indicate degrees of social proximity. Overall this suggests a move from formalised distance towards a sort of formalised friendliness.

Decline of pre-performatives and the emergence of the secretarial 'we'

Two further trends are identified by the quantitative analyses. First of all there is something of a decline in the first person pronoun '*I'* and an overall increase in the first person plural pronoun *'we'.* When examined in closer detail, this appears to be related to the decline in pre-performative phrases such as '*I am directed to', 'I am to',* and *'I beg to'*. Where previously secretarial authors performed their role as conveyor of the message, as in example (1),

(1) *"I am directed by the Postmaster General to acknowledge the receipt of your letter of the 7th ultimo..."* (1870_04_02_FIS_DHC))

Increasingly authors made the distinction between themselves as the conveyor of the message and the corporate body on whose behalf they write with personal pronouns (i.e. '*I'* as conveyor of the message, *'we'* as source of the message).

This decline in pre-performative clusters occurs in the BTCC around the same time as the decline in deferential formulas such as '*your obedient servant*'. The overall effect of these changes is that we end up with a business language that is more democratised in the sense that the corporate and personal hierarchies are not explicitly performed. However, this democratisation also seems to have contributed to the impression that business language has become more impersonal, and that there has been a shift 'from the individual to the corporate dimension of letter-writing' (Del Lungo Camiciotti, 2006:171).

Research on the corpus is ongoing, with the inclusion of additional material planned for late 2017. For now the BTCC has started to fill the gap in historical business correspondence data for this period, and provided a new way of engaging with the rich historical material in the BT Archive.

References

Brown, P. and Levinson, S. (1987) *Politeness: some universals in language usage*. Cambridge: Cambridge University Press.

Del Lungo Camiciotti, G. (2006) '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

Scott, M., (2012) *WordSmith Tools version 6*, Stroud, Lexical Analysis Software

Stubbs, M (2007) 'On texts, corpora and models of language' in Teubert, W and Mahlberg, M. (ed.) *Text, Discourse and Corpora: Theory and Analysis*. London, Continuum, 127-162