Denison, David. 1994. A corpus of late Modern English prose. In Merja Kytö, Matti Rissanen & Susan Wright (eds.), *Corpora across the centuries: Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, 25-27 March 1993* (Language and Computers - Studies in Practical Linguistics 11), 7-16. Amsterdam and Atlanta GA: Rodopi.

A corpus of late Modern English prose

David Denison
University of Manchester

1. Purpose

A corpus of late Modern English prose has been put together for a very specific purpose: to help with a chapter, Denison (in prep.), on the syntax of English over the last two centuries. Amongst other things, the corpus is intended to bridge the gap between the Helsinki Corpus and the various corpora of Present-day English so that measurements of frequency can be made, since most syntactic change in the period has been statistical rather than categorical. It would be pleasing if others could make use of it too, and subject to any copyright limitations (see §3.1 below), I would be happy to make it available.

The research fund of the Faculty of Arts, University of Manchester, covered the cost of employing an undergraduate student, Graeme Trousdale, to scan and code the corpus. Much helpful advice was given by Merja Kytö of the University of Helsinki.

2. Selection

Time and resources were very limited, so it was not possible to have a wide range of genres. The category ‘informal letters’ was selected as one that would represent ordinary language reasonably well, and which was available in the large corpora for purposes of comparison. The careful preparation of Nevalainen (1991: 109-11) was plundered for the titles of five suitable published collections of letters, which are listed under ‘Texts’ in the references below. It turned out that one of the collections, Amberley, had diary entries and letters intermingled. Both were included.

2.1. Size and sampling

The corpus stands at about 100,000 words of text, 105,000 words altogether. It currently occupies 555 KBytes on disk.

I decided to go for contiguous blocks of material, and the approximate number of pages needed to provide 20,000 words was worked out for each of the editions. Then for each one the starting page was determined randomly.[1] The effect, it so happens, was to produce a tolerably even spread of letters as far as date of writing is concerned:

Table 1: Distribution by date

DECADE	1860-69	1870-79	1880-89	1890-99	1900-1909	1910-19
Approx. WORDS (thousands)	20	20	6	13	20	20

Table 1 is based on approximate counts done by Graeme Trousdale. No decade in the overall range 1860-1920 is wholly unrepresented, though 1880-89 is somewhat low. (Later the material was grouped into two periods of thirty years each.) The spread based on date of birth of the writer is rather narrower: Lord and Lady Amberley (the main writers in that collection) were born in 1842, Bell in 1868, Dowson in 1867, Green in 1837, and the Webbs in 1859 and 1858.

2.2. Text entry

The chosen texts were scanned in on an HP Scanjet connected to a PC running OmniPage software. Files were converted to Ascii and rearranged so that no letter was divided across a file boundary. The scanner performed worst on the print with many ligatures, but even here its success rate was quite acceptable. Obvious mistakes were corrected at once, and printouts were then proofread and the corrections entered. Before distribution they will be proofread again against the original books.

2.3. Coding

As far as possible, coding of the material was done on the pattern established for the Helsinki Corpus, for the convenience of users and to facilitate comparison.

3. Problems and questions

I mention here the types of problem which have arisen, some of them not yet resolved. The section is organised along the lines of Kytö (1991: Part One).

3.1. Form of text files

3.1.1. Character set

Unlike the Helsinki team, whose practice was based on the then requirements of mainframe computers, I decided that 8-bit Ascii was preferable to 7-bit, as it is practical for all PCs and simplifies coding of nearly all the familiar European accented letters.

3.1.2. Lineation

I wished to preserve lineation of the original printed editions for ease of reference. This raises a problem with long lines, whether caused by a compact original typeface or by editorial insertion of comments and codes. The Dowson letters quite often, and the Webb letters occasionally, have line lengths greater than 80 characters; there are 44 such lines. Two solutions offered themselves:

(A) to allow lines longer than 80 characters

(B) to ensure that lines longer than 65 characters have a space and the # 'line continues' character at columns 64 and 65, with the remainder of the line carried over

Solution (A) is simpler to implement and quite practical for most text readers and editors. I think that 95 columns would be enough for the longest. Solution (B) would conform to Helsinki practice – see Kytö (1991: 23). For the time being a (B)-like solution has been adopted, though the # character is not in column 65. (For convenience in this report, the lineation of the corpus is ignored in cited examples.)

3.2. Coding

I decided not to attempt to follow SGML/TEI conventions, since I and other likely users would probably wish to use this corpus as a supplement to the Helsinki Corpus. Accordingly the text is coded à la Helsinki, with COCOA-type codes within caret brackets, one per line. These codes are mostly bunched at the beginning of each block of letters with a given writer and recipient, apart from page number codes. There are also text-level codes embedded in the text.

3.3. Text-level codes

3.3.1. Foreign language

The Helsinki conventions for foreign language material are to enclose them as follows (Kytö 1991: 30):

(1) Thought her agreeable but very foreign. Almost the moment after her arrival she told me that her aunt had told her it would be dangerous to know such men as me because of my (\"esprit"\) & the (\"coeur"\) shown in my article on War. I was much amused.

(1873 Amberley p.537)

In attempting to follow this convention we have found many tricky cases. Are *adieu, ennui* to be bracketed as French words? (The former hesitantly Yes, the latter No, except in a French context.) Is *ménage* foreign? (Yes, because of the acute accent.) Are *eheu, seriatim, Verb Sap* Latin? (The first two, Yes.) Is *Fräulein* German? (Yes.) Some of the writers, Dowson particularly, interlard their letters with snatches of genuine French, Italian, German, Latin and Greek, as well as playful neologisms and spellings like:

(2) Don't forget that we must try & work a (\theatrum\) this week.

(1890 Dowson 91)

(3) Shocquing!

(1889 Dowson 72)

(4) could you then five-o'clocquer with me at the Arts & Letters at 5.30.

(1890 Dowson 94)

There are bound to be inconsistencies in the decisions taken.

Passages of Greek are neither reproduced nor transliterated. Their presence is signalled by inserting:

(5) [^GREEK CHARACTERS^]

in the text, using the Helsinki conventions for 'our comment' (Kytö 1991: 35). Short passages in languages which use Roman script have been left in, but bracketed as in example (1).

3.3.2. Italicisation

Italicisation, presumably representing the writers' original handwritten underlining,[2] has not been preserved in the Ascii files and has not been coded (though the scanner was able to distinguish it).

3.3.3. Misspellings

There are a few children's letters in Amberley, but even among the adult letters – all by educated men and women – occasional misspellings crop up, for example *accomodation, guage, its* [= *it is*]. Naturally these are left uncorrected in the corpus. Rather than insert editorial comments each time, it might be worth providing a list for the benefit of users interested in lexis rather than spelling.

3.3.4. Abbreviations

In the same spirit I have not expanded such abbreviations as *f. c'd* [= *fair copied*], *shd* [= *should*], *Yrs* [= *Yours*], as I intend rather to provide a list. As Raymond Hickey pointed out at the Colloquium, in both cases it would be easy for users with the LEXA suite of programs to make a look-up table for expanding automatically those abbreviations and misspellings which do not correspond to ordinary spellings; however, note that the original text could not afterwards be restored from the corrected version by reversing the process.

The abbreviation M^{me} [= *Madame*] occurred once with superscript:

(6) I have just completed fair copying Chap II (new) of M=me= de V.

(1890 Dowson 92)

The coding in (6) follows Kytö (1991: 27). (Note that the equals symbol is used once elsewhere as part of the text.) Another abbreviation not in the Ascii set, ∴ [= *therefore*], was

coded as follows:

(7) I have nothing on earth to tell you therefore [^EDITION: ABBREV. AS 3 DOTS^] I may as well set to & write you a lengthy epistle.

(1890 Dowson 82)

3.3.5. Cancelled text

Text which is indicated in the editions as having been struck through and cancelled by the letter-writer is shown as follows:

Come back though soon to your old habitations even to the [^CANC. TEXT: square^] fields of Mesech, & let us Haymarket & Gaietieize & and Polandize to the end of our days.

(1889 Dowson 78)

3.4. Reference codes

3.4.1. Identification of text samples

At the time of writing the final layout of the files has not been decided, so the following sample represents provisional codings. Each file begins with a series of reference codes like this:

(8)
<B CLBELL>
<Q L89 XX CORP BELL>
<N LET TO H.B.>
<A BELL GERTRUDE>
<C L89>
<O 1890-1919>
<M X>
<K X>
<D ENGLISH>
<V PROSE>
<T LET PRIV>
<G X>
<F X>
<W WRITTEN>
<X FEMALE>
<Y 40-60>
<H PROF>
<U X>
<E INT UP>
<J INTERACTIVE>
<I INFORMAL>
<Z X>
<S SAMPLE X>

See Kytö (1991: 40-56) for details of the coding system. Full details of the edition used are then given in an 'our comment' bracket. As long as writer and recipient are held constant, only reference codes for the page number are updated. When writer and/or recipient change, then the full list as in (8) is repeated with salient changes. Within the sample list given above

I simply note that 'L89' in the Q and C fields refers to a corpus called 'L' (for 'Late Modern English'), subsection '89' (that is, beginning in the 1890s).

3.4.2. Social status and relationship

The Helsinki conventions provide for elaborate coding of the sex, age and social rank of writers and of the participant relationship between writer and recipient. Where the information was conveniently available we have included it, as can be seen in (8) above, but frequently it was not practical to do so within the constraints of time, and a non-committal 'X' has been used. And Graeme and I have not always found it easy to interpret the Helsinki conventions of social rank in relation to our nineteenth- and twentieth-century writers.

4. Future plans

4.1. Dissemination

I remain unclear as to the copyright status of the material and on limitations on usage beyond my own private study and research. As can be seen from the publication dates, some at least of the printed editions I used are clearly still within copyright. I plan to check with the publishers concerned. No decision has been taken on whether to index the files with WordCruncher or some other concordance program(s).

4.2. Extension of the corpus

It is possible that further material will be added: certainly it would be useful for my own research to have a larger corpus. Ideally one would widen the range of dates, add other genres such as dialogue from comedies, in fact ultimately – though this is well beyond my own ambitions – aim for similar coverage to that found in the Helsinki Corpus. But I have no immediate plans.

Notes

1 To Graeme's weary disbelief, I noted the available range of starting pages which would allow a 20,000-word sample to be taken, then insisted on repeated tossing of a coin to halve the range. So much for high tech.

2 But such cases of italics as '(*Pace* the small impudent man, ...)' (1862 Green p.108) may be editorial.

References

(1) Texts

Amberley = Russell, Bertrand, and Patricia Russell (eds.). 1937.
The Amberley papers: The letters and diaries of Lord and Lady Amberley, vol. 2, 512-51. London: Leonard & Virginia Woolf at the Hogarth Press.

Bell = Bell, Lady (ed.). 1927.
The letters of Gertrude Bell, 2 vols. Vol. 1, 396-403; vol. 2, 404-55. London: Ernest Benn.

Dowson = Flower, Desmond, and Henry Maas (eds.). 1967.
The letters of Ernest Dowson. 110-59. London: Cassell.

Green = Stephen, Leslie (ed.). 1901.
Letters of John Richard Green. 72-123. London: Macmillan.

Webb = Mackenzie, Norman (ed.). 1978.
The letters of Sidney and Beatrice Webb, vol. 1, *Apprenticeships 1873-1892*. 270-319. Cambridge: Cambridge University Press in co-operation with the London School of Economics and Political Science.

(2) Other works cited

Denison, David. in prep.
Syntax. In *The Cambridge history of the English language*, vol. 4, *1776-present*, ed. by Suzanne Romaine. Cambridge: Cambridge University Press.

Kytö, Merja. 1991.
Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts. Helsinki: Helsinki University Press for Department of English, University of Helsinki.

Nevalainen, Terttu. 1991.
BUT, ONLY, JUST: Focusing adverbial change in Modern English 1500-1900. Mémoires de la Société Néophilologique de Helsinki 51. Helsinki: Société Néophilologique.