Mark Faulkner
Description of MANCASS C11DB for Medieval Big Data Colloquium, Dublin, June 2017

MANCASS C11 Database

The MANCASS C11 database, developed under the leadership of Professor Donald Scragg, with funding from the Arts and Humanities Research Council, launched in 2004. The database had been unavailable for several years when Mark Faulkner, then of the University of Sheffield, contacted Professor Scragg in October 2014. With Professor Scragg's help, he recovered the original data from Manchester in March 2015. Since then, it has been in his possession, though not publicly available.

Given that Faulkner was not involved in the original project, he has no direct knowledge of it. The following description is therefore based on published reports, personal communications with Professor Scragg and his inspection of the surviving data.

Its scope and size

According to the original Palaeographical Introduction, which survives as a PDF document among the materials saved from Manchester, the Database inventoried, on the basis of the catalogues and handlists of Ker, Sawyer and Pelteret, 165 manuscript books, 70 documents, and 14 continuous glosses that contained written English from the eleventh century. Within these 249 sources, 1800 texts were identified. According to Scragg 2009, writing when further expansions had been made, over two thousand texts had been identified, of which over half were available in the database, collectively constituting over 1 million words.

Pleasingly, what survives is in significant ways more extensive than the resource Scragg 2009 describes: a database of 282 manuscripts and documents and 1883 texts, along with transcriptions of 978 of them, ranging of the 27,000 words of the D-version of the Anglo-Saxon Chronicle in Cotton Tiberius B. iv to a one-word rubric in CCCC 422. Collectively there are 1.8 million words; the average text is thus just over 1800 words in length.

Additionally, there survives at least a significant portion of the data from a follow-up project, 'English Glosses in Eleventh Century Anglo-Saxon Manuscripts', AHRC-funded from September 2007 to August 2009. According to Powell (2010), the project had identified 85 manuscripts as containing relevant material, of which at the time of writing she had viewed 36, recording 5600 marginal and interlinear entries.

Data survives from 48 manuscripts, principally those in Cambridge and London, though six Bodleian manuscripts are also represented. No list of the other manuscripts intended for inclusion survives, though Powell does say it was based on published sources, principally Ker's *Catalogue*, so it should not be impossible to reconstruct. According to Powell (2010), staffing and technical issues prevented the inclusion of the glosses in the database proper; they are therefore recorded in a series of PDF documents where the annotations are classified and tabulated with their contexts. A typical entry is below:

CCCC 178

Ker Item #	Page #	Line #	Type	Main Text	Gloss	Location on Page	Location in Word/Line	Hand	Cameron Number	Sense Unit
A1	1	26	alteration	Mar	an	above line	finally	main	B1.1.2	[000800 (178.14)] **Maran** kyððe habbað englas to gode þonne menn. & þeahhwæðere hi ne magon fulfremedlice understandan ymbe god.

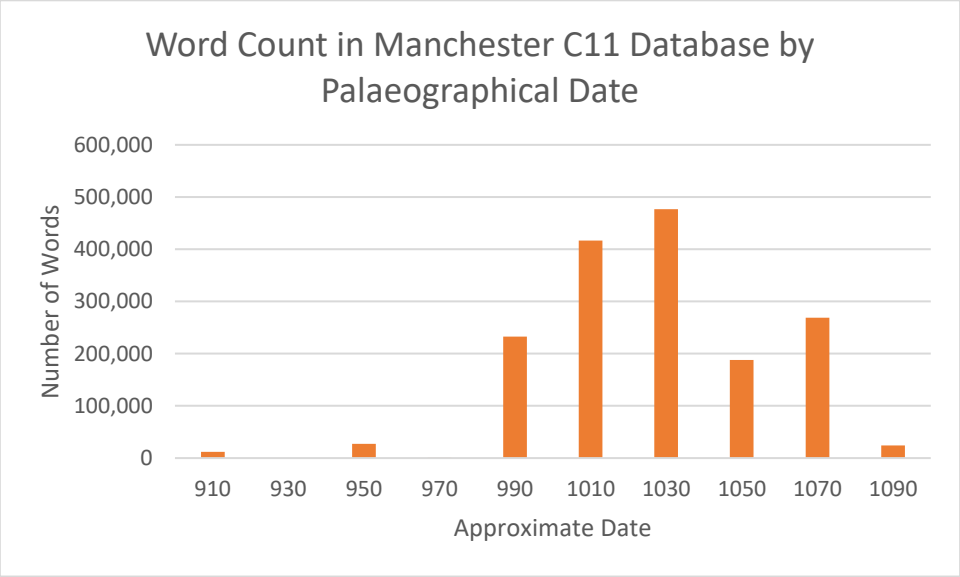
These tables range in length from the 577-page record of annotations in CCCC 162 to a single-page record of glosses in CCCC 391. The average file size is 201kB, suggesting an average table of around 50 pages. Given changes of policy regarding the amount of context provided and the format in which

the tables are preserved, it is difficult to obtain an accurate estimate of how many glosses this comprises, but it is a lot.

The criteria for inclusion: date, text type, dialect etc.

As the project documentation points out, what distinguishes the Manchester C11 database from the Dictionary of Old English Corpus is that it contains transcriptions of multiple texts of the same work. Thus, for example, the database contains transcriptions of nine separate copies of Ælfric's *De initio creaturae*, that is, all the surviving copies except the three of twelfth-century date and the eleventh-century but 3cm-wide fragment in Brasenose College Oxford.

The terminal dates for the project appear to have been 980x1099, with the vast majority of the texts necessarily dated paleographically.[1] I have not encountered an explicit statement of how the 978 texts included were chosen from the 1800 or so identified as meeting the criteria for inclusion; I suspect pragmatic grounds were key. Overall, as the graph below shows, coverage of the eleventh century is slightly uneven, but this may very well reflect shifts in the frequency of manuscript production (an interpretation that could be tested by a day's number-crunching).



Turning to how representative the Database is of different centres of manuscript production, we note that only about a third of the texts in the Database are localised and several centres loom particularly large:

Production Location	Texts	Word Count
Abingdon	9	2,157
Bath	5	33,660
Bury St. Edmunds	15	47,881
Canterbury, Christ Church	41	56,217
Canterbury, St.Augustine's	1	352
Durham	1	58
Exeter	48	134,203
Exeter[2]	5	1,839
Glastonbury	1	3,048

[1] The surviving database contains a handful of material (comprising just over 2% of the total) there given the palaeographical dates of s. x^{in}, x^{med} and x^{2}. At the moment, it is unclear why.
[2] It is unclear to me why two separate location codes were used for Exeter. Is one intended to denote the monastery, the other the cathedral?

London, St. Paul's	25	16,206
Malmesbury	2	1,551
Rochester	2	683
Salisbury	1	4
Sherborne	35	5,453
Winchester, New Minster	37	26,935
Winchester, Old Minster	4	1,137
Worcester	94	151,604
York	8	2,622
TOTAL	334	485,610

The only genre classification used appears to have been prose versus verse. This is coded in the XML header of each text, but is for the reasons explained below not at present accessible. These headers also coded dialect, so it will in the future be possible to assess the database's representativeness in this respect as well.

The extent to which texts are marked up: if they are tagged, how are they tagged? Details of the format in which marked-up texts are encoded, and of the extent to which the files are available for independent use / reuse

According to Powell 2004, witnesses were originally transcribed as text files, then marked up in XML. The transcriptions prioritised recording what the original scribes wrote, ignoring any later corrections. The XML encoding appears to have involved the addition of a header, of the following format:

```
<cota crdate="940610" update="940610" id="xxx">
<header><fileDesc><titlStmt>
<cameron> b1 9 1 </cameron>
<shtTitle> ægram </shtTitle>
<title>aelfric grammar part</title>
</titlStmt>
<extent>587</extent>
<srcDesc><biblRef> ms durham b iii 32 zupitza 1880 1-296 </biblRef></srcDesc></fileDesc>
<profdesc><txtclass><catref target="pls">c=prose p=late d=saxon</catref></txtclass></profdesc>
</header>
```

Only one text, that reproduced above, survives in as a freestanding XML file. The XML for the other texts survives embedded in a Sql database.

According to the introductory comments published when the website launched in 2004, 'the project team tried very hard for a very long time to automate this part of the process [*scil*. lemmatisation], but it proved impossible'. The files were therefore manually lemmatised by Scragg and Powell. It should be noted however that this lemmatisation was not recorded in the xml texts (or if it was, this aspect of it does it survive); rather, it was used to populate an exhaustive list of lemma groups with all the spellings of each lemma attested in the database. Thus, lemma group 25, representing the OE antecedent of PDE *king*, comprises 132 separate forms of the noun, which represent 50 different spellings of the stem (e. g. *cyncg-*, *kync-*, *kyninc-* etc).

The XML files were however processed in two further ways. They were divided into numbered segments, which, where possible, accord with the divisions used in the Dictionary of Old English Corpus. Second, words in Latin and passages where the original reading could not be recovered due to later correction or damage were assigned the tag <ignore/>. A typical line of transcription from one of the XML files therefore looks something like this:

<s><lineRef>xxx0019</lineRef><biblRef>4 18</biblRef><ignore>littera</ignore><content> is stæf on englisc and is se læsta dæl on bocum and untodæledlic </content></s>

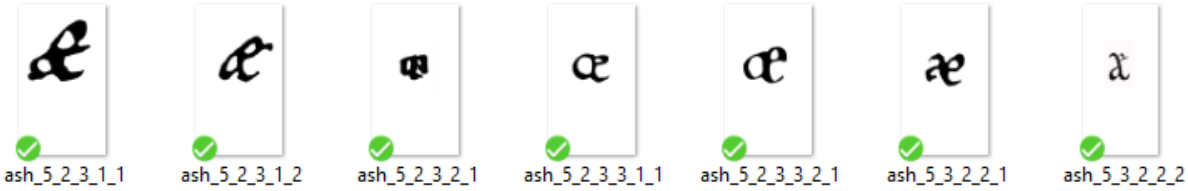
Confronted with the Sql database and in the absence of able friends or funding forced to fall back on his own industry, Faulkner found a way to extract the text of each item from the database as a .txt file. The C11 spellings database is therefore now available as a set of 978 .txt files, which can be searched using third-party software of the user's choice (e. g. AntConc). The extensive metadata (see below) collected in the database is available in large (20 Mb) Excel spreadsheet.

Though this state of affairs is better than that which prevailed before the data was recovered, it is still far from ideal. The database is not publicly available; Faulkner's ad hoc converted .txt files omit anything between <ignore/> tags from the transcriptions (which is fine for many kinds of linguistic research but hinders use of the transcriptions as manuscript surrogates); and, while all the information from the Database can be accessed through the spreadsheet, piecing together the details of a particular text involves flicking from large sheet to large sheet within the spreadsheet and is often extremely laborious. In short, it would be of major benefit to get the database back online.

The extent of metadata relating to texts: what information available about text date, authorship, localisation, transmission, editorial history etc?

The MANCASS database was a major exercise in linking data. It contains a list of all eleventh-century manuscripts and documents containing English, and, linked to this, a list of every text these manuscripts and documents contain. Each work is identified by its Cameron number. The database also collected extensive data on the manuscripts and documents, enumerating and dating every scribal stint each contains, identifying scribes who wrote more than one stint, in addition to offering a brief palaeographical description and, where appropriate, a list of published facsimiles of each hand. All of this survives, but, because it is currently accessible in an Excel document rather than an Sql database, is cumbersome to use.

A major component of the database as it was originally envisaged were palaeographical 'identikits' of the various scribes inventoried. According to the 'Palaeographical Introduction' from the original website, eleven minuscule letters (a, æ, c, d, e, g, h, s, þ, ð, y), ten majuscule letters (A, Æ, D, E, G, H, M, N, T, Ð), one abbreviation (7, 'and') and the treatment of ascenders were chosen for study. There survive images of 180 exemplary forms of these different letters, for instance these seven forms of æsc:



However, as far as I can see there is nothing in the surviving materials that records which letter-forms were found in the stints of which scribes. Had they survived, these analyses would have offered significant evidence with which to evaluate the validity of the palaeographic dates traditionally assigned to the various texts in the database.

The search protocols available

The fullest description of how the original database could be used is Powell (2004). According to her account, the main means of interrogating it was by searching for words or stems. Searching for the

word 'cyning' told the user the word occurred 812 times in the database, in 45 different forms which comprised 21 different stems (these numbers differ from those given above because the data in the database was significantly augmented after Powell wrote). The user could then click on any of the hits and see the form in context; clicking in turn on this contextualised reading would bring up the relevant readings from other texts of the same work, if any survived and were present in the database. Early documentation advised users to note that the database could not be used for searching for spellings of short, common words (e. g. *and*).

According to Powell (2005), a number of other search functions were available, including options to track spelling variants involving geminates and single-letter alternations in the stems of words (e. g. all words where *y* is sometimes substituted for *i*). The ability to search for spelling frequency by palaeographical era was in development.

These various options must have made for a tool of considerable power, but it is worth noting two weaknesses, one major, one minor. The minor deficiency is noted by Scragg (2009), which is that the database only recognised as variant texts of the same work texts which have the same Cameron number. Thus, the texts of Vercelli 2, Vercelli 21 and Napier 40 do not show up as separate witnesses to the same work, even though they must be in some way related. The more major issue, mentioned by Powell (2005), is that the database could not disambiguate homographs (presumably because the lemmatisation was not encoded in the XML). Thus, searching it for *witan* would return forms of *witan*, 'depart' and *witan*, 'know', with the user needing to manually distinguish them by looking at each hit in context. Obviously, this would have made the MANCASS database significantly less helpful for particular kinds of linguistic study than corpora like LAEME that are lexically tagged.

As mentioned above, the transcriptions made for the MANCASS database currently exist as a series of .txt files. These can be interrogated by any method of search possible with existing software (wildcard, regex etc).

A list of specimen research projects that have used the text archive

Scragg (2003) [not seen]; Scragg (2006); Scragg (2006b) [not seen]; Scragg (2009); Scragg (2012).

Details of envisaged future developments

Getting it back online in some form!

References

Kathryn Powell, 'The MANCASS C11 Database: a tool for studying script and spelling in the eleventh century', *Old English Newsletter* 38.1 (2004), 29-34.

Kathryn Powell, 'The English Glosses in Eleventh-Century Anglo-Saxon Manuscripts', *The Heroic Age* 13 (2010).

Alexander Rumble, 'The Palaeographical Material in the C11 Database', *MANCASS C11 Database* (published online 2004).

Donald Scragg, 'Standard Old English and the Study of English in the Eleventh Century', *Old English Newsletter* 35.1 (2001), 24-26.

Donald Scragg, 'Standard Old English: Scribal Practices in the Eleventh Century', *Revista Canaria de Estudios Ingleses* 47 (2003), 37-44 [not seen].

Donald Scragg, 'Ælfric's Scribes', *Leeds Studies in English*, n.s. 37 (2006), 179-89.

Donald Scragg, 'Rewriting Eleventh-Century English Grammar and the Editing of Texts', *Bells Chiming From the Past: Cultural and Linguistic Studies on Early English* eds. Isabel Moskowich-Spiegel and Begoña Crespo-García. (Amsterdam, 2006), 195–208 [not seen].

Donald Scragg, 'Studies in the Language of the Copyists of the Vercelli Homilies', in *New Readings in the Vercelli Book* ed. Samantha Zacher and Andy Orchard (Toronto, 2009), 41-61.

Donald Scragg, 'Sin and Laughter in Late Anglo-Saxon England: the case of Old English *(h)leahtor*', in *Saints and Scholars: new perspectives on Anglo-Saxon literature and culture in honour of Hugh Magennis* ed. Stuart McWilliams (Woodbridge, 2012), 213-23.