Metaphor in End-of-Life Care:
UK Data Service deposit documentation

This file documents the structure and contents of this data deposit, which contains the data sources generated by the ESRC-funded "Metaphor in end-of-life care (MELC)" project (grant reference: ES/J007927/1).

For more information about MELC, see http://ucrel.lancs.ac.uk/melc/ (the project website includes references to the project's published papers, which contain more detail on how we collected and analysed our data than there is space for here).

If you use this data, please credit the MELC project in any resulting publication by using (1) the DOI of this data deposit plus either (2) our project website or (3) one of our publications describing the data collection.

What this deposit includes

This deposit contains five different *text corpora*, that is, very large collections of language data. One of these consists of transcribed *interview data*. The other four consist of *online data* drawn from publicly-viewable sources on the World Wide Web.

The deposit does not include any data sources that we used on the project that we did not create, but rather adopted from earlier research. Our publications contain full reference details for such data where relevant.

Anonymisation

The interview data has been fully anonymised; all instances of a confidential personal name in the text have been replaced by the placeholder code **_NAME_**, and all instances of a placename that might allow a person to be identified through their place of work have been replaced by the code **_PLACE_**. In frequent cases, entire utterances are omitted because the identifying information was more extensive than could be effectively anonymised by masking single words.

Names of public figures and of places that are not related to the informant have not been anonymised.

The online data has not been anonymised, because its contents are already publicly accessible via the WWW and therefore are not confidential.

File formats

We use two formats: XML (for data collected from online sources) and plain text (for interviews, where there is no file-internal structure that would require XML encoding).

The XML files do not use a DTD; the tag schema is very simple and self-documenting for the most part (some brief comments on the XML used is noted for each corpus below).

All files use UTF-8 character encoding (*without* a leading byte-order mark character) and Unix-style line breaks.

Organisation of the deposit

There are 5 separate files within this deposit:

- This file documents the data.
- Each of the other files is a compressed archive (.zip) containing a single folder or shallow folder hierarchy, within which there is a set of plain-text or XML files.
- Each of the folders within the archives contains one of the five corpora listed below.

The five corpora

Transcribed interviews with healthcare professionals

This corpus is contained within the file **MELC-InterviewData.zip**.

It consists of 16 transcribed interviews with healthcare professionals. Each interview is contained within a separate plain-text file (*not* XML).

The total size of the data is approximately 481 Kb uncompressed.

The filename schema is as follows: an arbitrary unique letter as the ID code for the individual text, followed by an underscore, followed by the date of the interview (as *ddmmyy*).

The files are fully anonymised (see above).

Online data: posts from the "Macmillan" discussion forum

This corpus is contained within the file **MELC-OnlineData1.zip**.

It consists of a series of discussion threads from the online forum run by *Macmillian Cancer Support* (available, at time of writing, at http://community.macmillan.org.uk/) and primarily utilised by persons who have, or have had, cancer, or who are caring or have previously cared for such a person. The data was collected in November 2012 via automatic web spidering.

Each file contains a single discussion thread (delimited by <thread> tags).

There are 30,221 separate files. The filename of each is the unique XML identifier of the <thread> element within the file. These consist of the mnemonic code "mm12" (which

abbreviates “Macmillan data collected in 2012”), followed by an underscore, followed by an arbitrary decimal number.

Within each thread posts are given in chronological order (and delimited by <post> tags). The author of each post is given in the XML markup, using the online handle used on the Macmillan forum system (frequently pseudonymous). As all the text, as well as the metadata, was collected from openly-accessible webpages that did not require a login to download, we have not anonymised the data.

Within each <post>, XML
 tags indicate line or paragraph breaks in the original forum post. None of the other formatting of the original post has been preserved.

The total size of the data is approximately 358 Mb when decompressed.

Online data: posts from the “doc2doc” discussion forum

This corpus is contained within the file **MELC-OnlineData2.zip**.

It consists of a series of discussion threads from the online forum “doc2doc”, run by the *BMJ Publishing Group* (available, at time of writing, at http://doc2doc.bmj.com/) and primarily utilised by healthcare professionals. The data was collected in March 2013 via automatic web spidering. As well as discussion forum contributions, data has been collected from blog post comments.

The XML format, file naming convention and internal file structure is identical to the “Macmillan” corpus, discussed above.

There are 7,662 files.

The unique identifiers of the threads/files consist of the mnemonic code “doc2doc”, followed by an underscore, followed by an arbitrary hexadecimal number.

The total size of the data is approximately 58 Mb when decompressed.

Online data: manually-sampled posts by healthcare professionals (1)

This corpus is contained within the file **MELC-OnlineData3.zip**, subfolder **bmjcr**.

It consists of healthcare professionals’ responses to *BMJ* comment articles (in the “Comment and Response” section, thus the mnemonic *bmjcr*. The articles were accessed via www.bmj.com and were identified manually as being relevant to the topic of the MELC project’s research.

The individual posts are marked up using <post> tags. Some files contain more than one post. However, since the collection of posts in each file may not be a single full sequential thread, the containing element is instead <postGroup>.

There are 20 files. Filenames are consist of the mnemonic code “bmjcr”, followed by an underscore, followed by the date of the post (as *ddmmyy*). Where this leads to more than one file having the same name, an additional letter is added (*a*, *b*, *c* …)

The total size of the data is approximately 140 Kb when decompressed.

Online data: manually-sampled posts by healthcare professionals (2)

This corpus is contained within the file **MELC-OnlineData3.zip**, subfolder **bmjblog**.

It consists of blog posts from http://blogs.bmj.com. The blog posts were manually identified as relevant to the concerns of the MELC project.

The XML formatting is as in the *bmjcr* corpus (see above) except that there are no <postGroup> elements since every file is a single <post>.

As the post metadata was coded manually rather than automatically, it is grouped within the *title* attribute, and there are no automatically generated arbitrary IDs; instead, each filename contains the mnemonic code “bmjblog”, followed by an underscore, and then the name of the author and the date of the blog post (which together are unique).

There are 44 files.

The total size of the data is approximately 215 Kb when decompressed.