Email in the Australian National Corpus

Andrew Lampert

CSIRO ICT Centre, North Ryde, Australia and Macquarie University, Australia

1. Introduction

Email is not only a distinctive and important text type but one that touches the lives of most Australians. In 2008, 79.4% of Australians used the internet, ahead of all Asia Pacific countries except New Zealand (Organisation for Economic Co-operation and Development, 2008). A recent Nielsen study suggests that almost 98% of Australian internet users have sent or received email messages in the past 4 weeks (Australian Communications and Media Authority, 2008), making email the most used application on the internet by a significant margin. It seems logical to embrace a communication medium used by the vast majority of Australians when considering the text types and genres that should be included in the Australian National Corpus.

Existing corpora such as the British National Corpus (2007) and the American National Corpus (Macleod, Ide, & Grishman, 2000) provide many insights and lessons for the creation and curation of the Australian National Corpus. Like many existing corpora, the British National Corpus and the American National Corpus contain language data drawn from a wide variety of text types and genres, including telephone dialogue, novels, letters, transcribed face-to-face dialogue, technical books, newspapers, web logs, travel guides, magazines, reports, journals, and web data. Notably absent from this list are email messages.

In many respects, email has replaced more traditional forms of communication such as letters and memoranda, yet this reality is not reflected in existing corpora. This lack of email text is a significant gap in existing corpus resources. The Australian National Corpus has an opportunity to fill this gap by including email data in the scope of texts gathered, and to provide a differentiated and valuable corpus resource for researchers across a wide variety of disciplines.

2. The Need for Email

The lack of real-world collections of email messages is a major impediment that continues to hinder research progress in a variety of fields exploring language use, social interaction, information processing, and other aspects of how people use email to communicate.

2.1. An Example: The Enron Dataset

The single notable exception to this lack of real-world email data has been the release of a large collection of email messages during the Federal Energy Regulatory Commission (FERC) investigation into the collapse of Enron in the United States. The release of almost 500,000 email messages during the legal proceedings led to the creation of the Enron email dataset (Klimt & Yang, 2004). The release of the Enron dataset has been the single most significant advance in progressing email research, and the large volume of published research based on the Enron data (e.g., Dredze et al., 2008; Heer, 2005; Lampert, Dale, & Paris, 2008; McCallum, Corrada-Emmanuel, & Wang, 2004) provides compelling evidence of the value of real-world email datasets for researchers.

While the Enron corpus has undoubtedly provided a sound foundation for research grounded in observations of real-world email interactions, many researchers remain concerned that our insights are dominated to such a degree by the culture, norms, technology, and idiosyncrasies of a specific snapshot of a single company's patterns of email usage. There remains a significant need for further collections of real-world email data in order to conduct complementary and comparative research. Despite this

© 2009 Andrew Lampert. Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus, ed. Michael Haugh et al., 55-60. Somerville, MA: Cascadilla Proceedings Project.

compelling need, gaining access to other substantial collections of email messages has remained an elusive goal, and to date there has been no significant collection of Australian email data available for research purposes.

2.2. The Challenges: Copyright and Privacy

One factor that acts as a barrier against the release of email data to researchers relates to privacy concerns of email authors and recipients. Many people are understandably reluctant to make their sometimes intimate or personal email data available to researchers.

Compounding these potential privacy concerns, copyright restrictions are an even more significant impediment to the distribution of email data. In general, copyright ownership of the content in an email message rests with the author. When someone receives an email message, they do not own the copyright; it remains the property of the sender or the sender's employer. Attachments to email messages complicate matters further. The owner of copyright of an email attachment depends on the nature and author of the attachment, often different from the nature and author of the email message in which it is embedded. As a result, email attachments are often treated as works separate from any associated email messages.

The copyright situation is further complicated because many email messages include text not written by the sender of a message. This is common, due to both the norms of use across different email communities and the technical implementation of most email software that encourages selective quoting or wholesale inclusion of prior messages in a thread of conversation. At least theoretically, all authors who have contributed to an email thread would need to give permission for that thread of email messages to be released to researchers. This is because each contributor has copyright over a portion of the email text in the thread. This distribution of required permission across many people quickly makes the task of soliciting permission infeasible with any realistically sized collection of email messages.

A counter-position to the requirement for seeking permission from all contributors is to consider that an implied licence exists for certain uses of email content, for example when quoting previous text in an email response sent to a mailing list. The norms of most mailing lists would reasonably lead most people to expect any messages sent to the list to be quoted in response, and people contribute their email to the list on that basis. There are, however, many other uses for which such an implied licence would be difficult to establish. In particular, forwarding a received message to a third-party, let alone trying to gather a collection of email (even mailing list data) for inclusion in a linguistic corpus, is a use that would probably not be covered by an implied licence. Use in a corpus for academic research may be acceptable under fair dealing provisions for research or study that exist in the Australian Copyright Act. Usually, however, such fair dealing considers only individual research or study that makes use of small portions of a text, rather than an entire collection to be distributed and used by a community of researchers.

Thankfully, the situation is potentially more tractable for workplace email, where messages are written by employees using work resources and/or during work time. In such cases, copyright usually rests with the employer rather than the individual authors. To gather a collection of intra-company email sent and received by employees from a single workplace would thus require permission only from that single company, as the copyright owner. This is a potential avenue for the Australian National Corpus to gather a collection of Australian workplace email messages.

Despite the copyright and privacy challenges discussed above, we strongly believe that the benefits of including email in the Australian National Corpus outweigh the challenges of gathering such data. As we outline in Section 3, we may even be able to side-step many of these challenges through using an existing collection of Australian email messages to bootstrap the Australian National Corpus.

3. Collecting Email Data

Given the challenges posed by copyright ownership and privacy concerns, one of the biggest hurdles to including email in the Australian National Corpus is to determine how to gather a significant collection of real-world Australian email messages. Particularly for a nascent project like the Australian

National Corpus, this is no small undertaking. Thankfully, there may be a way to bootstrap the email component of the corpus through the efforts of an existing Australian email archive.

3.1. A Starting Point

During April and May 2008, the Powerhouse Museum and NineMSN ran a project called Email Australia that solicited personal contributions of email to create a public archive of Australian email (NineMSN & Powerhouse Museum, 2008). The project was inspired by a very similar project run by the British Library and Microsoft in the United Kingdom in 2007 (British Library, 2007).

A major motivation for both the British and Australian email projects was to publicise efforts to preserve a slice in time of the ephemeral messages we send to our friends, colleagues, and family, so that this information is not lost to future generations of historians, sociologists, linguists, and other researchers. These aspirations clearly complement the objectives of building the Australian National Corpus.

The Email Australia project reportedly received more than 10,000 messages from the Australian public. Submissions to the Email Australia archive were open to all Australians with an email address for a 6 week period in April and May 2008. Email was submitted to any of eight categories designed to solicit a cross-section of email communication. The eight categories were:

(a) Life and Laughter
(b) Touching Tales
(c) Family
(d) Love and romance
(e) Emails you regret sending
(f) Embarrassing typos
(g) Current affairs
(h) Complaints

To address copyright issues, participants had to warrant that they had obtained permission from all people who had authored material in their email message. So, for example, a forwarded email thread required permission from the authors of all quoted messages before it could be submitted. Only participants who warranted that permission had been obtained had their submissions accepted into the Email Australia archive.

Participants were also requested to ensure that identifying information such as surnames, email addresses, physical addresses, and phone numbers were either not present or were manually redacted from email messages before being submitted to the archive. Browsing a subset of approximately 1700 messages from the archive shows that not all participants paid close attention to this condition.

Participants were also requested to avoid submitting messages containing commercially sensitive information. Perhaps for this reason, none of the categories were targeted at soliciting workplace email – indeed the messages seen by the author are mostly of a personal nature. Other guidance given to participants was to avoid messages containing excessive profanity, defamatory content, third party intellectual property (e.g., poems and song lyrics), and material likely to cause offense or upset, or that any person mentioned in or contributing to the email content would not wish to be published.

Overall, the conditions attached to the submissions of email, including the requirement for permission from all relevant copyright owners, make the data collected in the Email Australia project a potentially useful starting dataset to be included in the Australian National Corpus.

3.2. Additional Email Data

In addition to bootstrapping the collection with email from the Email Australia project, the Australian National Corpus could follow the American National Corpus and solicit individual contributions of email data. This approach would allow individuals to contribute personal email data to the corpus, in much the same way as contributions were sought for the Email Australia project. Unfortunately, gaining a balanced and substantial corpus of email in this manner would be very difficult; the American National Corpus does not yet contain email text in any of their three released versions, despite the explicit request for email data on their contributions page.

Additionally, any email gathered in this manner would represent disconnected fragments of email from different contexts, lacking the coherence of something like the Enron dataset. Of course, the Email Australia data discussed in Section 3.1 also suffers from these same limitations.

As noted earlier, a possible avenue for a more substantial collection of email would be to gather a collection of intra-company email from a single workplace. This procedure would have the advantage of requiring permission only from the company, as the copyright owner. Gathering email from a former (i.e., failed or closed) company might also be a viable option.

A final avenue that might yield substantial collections of email data would be to look to various government organisations. Email messages created using Australian Government systems are Commonwealth records and must be managed in accordance with the Archives Act. They are also subject to related legislation such as the Freedom of Information Act, Privacy Act, and the Evidence Act, and may be subject to legal processes such as subpoena (National Archives of Australia, 2009). Many government email messages are stored in record management systems, and may under certain circumstances be accessible for corpus construction purposes.

4. Implications of Including Email

The inclusion of email in the Australian National Corpus requires some consideration to ensure that storage of corpus documents and associated collection processes accommodate rich conversational documents such as email messages.

4.1. Native Electronic Formats

One of the key requirements for maximizing the utility of corpus data, in particular for computational linguists, is to ensure that data is stored in a digital format that preserves as much as possible of the available information about the original context. For email data, this requirement means that wherever possible email messages should include more than just the body text and attached documents. In particular it is highly desirable to ensure full header information (i.e., sender and recipient information, subject line, dates etc.) is stored, along with any information about mailbox structure such as the folders into which messages are filed, and any labels, categories, or tags that have been applied to specific messages.

Attention must also be paid to storage of the actual content of email messages. Email messages are commonly authored in a variety of formats with varying degrees of sophistication. The simplest messages use plain ASCII or unicode text. Most email clients also support authoring messages in rich text, which allows text styling, fonts, and other presentation information to be specified. Beyond rich text, email messages are also commonly authored in HTML, which allows for the inclusion of images, graphics, sophisticated text styling, and layout control. In the Australian National Corpus, messages should be stored in their original format wherever possible to ensure that the presentation of each message remains as close to the author’s intended delivery as possible.

The Enron email dataset is a particularly useful dataset because the released data includes the full text of a large number of email messages in their original format, together with their associated full header information, and information about the different folders within approximately 150 different mailboxes into which the mail was organized by the original email users. This wealth of metadata has enabled researchers to use the Enron data for a variety of experiments in areas such as automatic message classification (e.g., Bekkerman, McCallum, & Huang, 2004), which would not be possible without this access to this information.

4.2. Implications for Corpus Design and Processes

Including email content in the Australian National Corpus is likely to have implications for both the corpus design and the data collection processes.

On the design side, the corpus must be able to handle potentially complex structures of messages in conversational threads. Additionally, email messages frequently contain attached documents in a huge variety of formats that can scale to several megabytes in size. The design of the corpus must be such that the corpus can store these and maintain information regarding the relationship between attached documents and associated email messages. As discussed in Section 4.1, the design of the corpus must also accommodate the retention of email messages in their native digital format including attached documents and any associated metadata, labels or tags, folder information, and message headers.

On the process side, thought needs to be given to the characteristics of email messages when designing constraints and requirements for submissions. The American National Corpus, for example, restricts submissions of individual documents to those containing no fewer than 2500 words (American National Corpus, 2007). Clearly such a limit is unlikely to be satisfied by almost all email messages. If email data is solicited, then thought must be given to appropriate and complementary submission criteria.

Additionally, the American National Corpus requires contributor(s) to *own* the copyright to all materials submitted, or that the materials must be in the public domain. Whilst clearly this is a justifiable requirement, it is, for example, more stringent than the terms and conditions used for eliciting data in the Email Australia project, which required participants to obtain the consent of people who own the copyright in any part of a submitted email message or attachment. More stringent submission criteria like those applied to the American National Corpus would likely prevent the inclusion of data from the Email Australia project in the Australian National Corpus.

5. Conclusions

Email has in many respects displaced several traditional forms of communication such as letters and memoranda and has become prevalent in the day-to-day lives of the majority of Australians. Despite the ubiquity of email as a text type, existing corpora of text largely omit email messages from their content.

While a number of issues, including copyright ownership and privacy concerns, make gathering email data for research purposes a challenging task, the value of such data to researchers across a wide variety of fields makes it worth addressing these challenges to ensure a representative sample of Australian email data is included in the Australian National Corpus.

To bootstrap efforts to gather Australian email data, the author believes that the data gathered by the Email Australia project, a partnership between the PowerHouse Musem and NineMSN, may be able to form a useful initial collection of email messages to be included in the Australian National Corpus.

6. Acknowledgements

The author would like to thank Matthew Connell, Computing and Mathematics Curator at the Powerhouse Museum, for providing access to examine some of the data collected during the Email Australia project. The author would also like to thank Cécile Paris and Robert Dale for their comments on an early draft of this paper.

References

Australian Communications and Media Authority. (2008, September). *Telecommunications today, report 6: Internet activity and content.* ACMA. Available at http://www.acma.gov.au/WEB/STANDARD/pc=PC_9058

American National Corpus. (2007). *Criteria for contributing documents to the ANC.* Available at: http://www.americannationalcorpus.org/authors.html

Bekkerman, Ron, Andrew McCallum, & Gary Huang. (2004). *Automatic categorization of email into folders: Benchmark experiments on Enron and SRI corpora* (CIIR Tech. Rep. No. IR-418). University of Massachusetts.

British National Corpus. (2007). *The British National Corpus, version 3 (BNC XML Edition).* Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available at http://www.natcorp.ox.ac.uk/

British Library. (2007) *First ever national email archive to be created.* Available at http://www.bl.uk/news/2007/pressrelease20070503.html

Dredze, Mark, Hanna Wallach, Danny Puller, Tova Brooks, Josh Carroll, Joshua Magarick, et al. (2008). Intelligent email: Aiding users with AI. NECTAR Paper. *Proceedings of American National Conference on Artificial Intelligence (AAAI), 23,* 1524-1527.

Heer, Jeff. (2005, June). Exploring Enron: A sketch of visual data mining of email. Paper presented at the Email Archive Visualization Workshop, University of Maryland..

Klimt, Bryan, & Yiming Yang. (2004). Introducing the Enron corpus. In *Proceedings of First Conference on Email and Anti-Spam (CEAS),* July 30-31, Mountain View, CA, USA. Available at http://www.ceas.cc/papers-2004/

Lampert, Andrew, Robert Dale, & Cecile Paris. (2008). Requests and commitments in email are more complex than you think: Eight reasons to be cautious. In David Powers & Nicola Stokes (Eds.), *Proceedings of Australasian Language Technology Workshop* (pp. 55-63). Hobart, December 8-10.

Macleod, Catherine, Nancy Ide, & Ralph Grishman. (2000). The American National Corpus: Standardized resources for American English. *Proceedings of the Second Language Resources and Evaluation Conference (LREC)* (pp. 831-836). 31 May – 2 June, Athens, Greece.

McCallum, Andrew, Andres Corrada-Emmanuel, & Xuerui Wang. (2004). *The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and academic email* (Tech.Rep.No. UM-CS-2004-096). Department of Computer Science, University of Massachusetts at Amherst.

National Archives of Australia. (2009). *Managing email. A new form of evidence.* Available at: http://www.naa.gov.au/records-management/systems/email/index.aspx

NineMSN & Powerhouse Museum. (2008). *Email Australia.* Available at http://emailaustralia.ninemsn.com.au/

Organisation for Economic Co-operation and Development. (2008). *Internet users, 2008.* Available from NSW Department of State and Regional Development at http://www.business.nsw.gov.au/aboutnsw/infrastructure/D10_internetusers_aust.htm

Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages

edited by Michael Haugh, Kate Burridge, Jean Mulder, and Pam Peters

Cascadilla Proceedings Project Somerville, MA 2009

Copyright information

ISBN 978-1-57473-435-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, e-mail: sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Lampert, Andrew. 2009. Email in the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 55-60. Somerville, MA: Cascadilla Proceedings Project.

or:

Lampert, Andrew. 2009. Email in the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus*, ed. Michael Haugh et al., 55-60. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2288.