Peer Review Comments

Article: Thiel, L, Sage, K and Conroy, P 2015 Normative Data for Email Writing by Native Speakers of British English. *Journal of Open Psychology Data* 3:e4, DOI: http://dx.doi.org/10.5334/jopd.aj

Article submitted: 19 December 2014

Article accepted: 15 April 2015

Article published: 22 May 2015

Editor: Jelte M. Wicherts

Responses for Version 1

Reviewer A: Jan Vanhove
Review Completed: 04 March 2015

This submission consists of a corpus of 3 ¡Á 42 short emails written by (I presume) native speakers of British English spanning a wide age range. In addition to the emails, which have been rendered anonymous and are untagged for parts-of-speech, morphology or syntactic structures, a table with each participant's sex, age and years of education is provided.

For the most part, the authors clearly describe how the data were collected (see below for some clarification requests). The selection of informants (Section 2) seems rather strongly skewed to the highly educated, but this will be obvious to re-users. The informant sample isn't strongly skewed towards young people, however, which is definitely an interesting feature of this corpus.

This corpus could definitely be useful to other researchers, but, as I outline below, I do have a couple of suggestions for enhancing its re-use potential.

Clarification requests and further suggestions:

1. Sample: I assumed that the speakers were all native speakers of British English (perhaps all recruited from the Greater Manchester area), but this isn't explicitly mentioned anywhere. Such information would be crucial for linguists and applied linguists. In fact, I'd recommend that you already explicitly mention this in the title and abstract (e.g. 'Normative data for email writing by native speakers of British English') so that this is clear from the outset. If they aren't all native speakers of BrE, it'd be useful to add this information to the table with the demographic variables.

2. Reuse potential: If 1. is clearly addressed, I could imagine that this corpus could be useful not only to researchers working on language disorders but also to researchers interested in, say, second language acquisition.

3. Data format: The corpus itself is stored in one docx-file whereas the three demographic variables for each participant are stored in another docx-file. Reusers, or people who are just skimming the corpus, don't have an easy automatisable way to link the texts to the demographic variables (i.e. other than manually looking it up). I wonder whether both sources wouldn't better be presented simultaneously, e.g. in a minimally tagged XML file. Something along the following lines perhaps (with some made-up meta-data):

```
<email>
 <emailID>ID1.1</emailID>
 <participant>ID1</participant>
 <sex>male</sex>
 <age>16</age>
 <education>12</education>
 <L1>BrEnglish</L1>
 <date>2011-03-09</date>
 <topic>meeting</topic>
 <salutation>Hi John,</salutation>
 <body>On this Friday etc.</body>
 <closing>William.</closing>
 <time>2m16s</time>
</email>
```

<email>
<emailID>ID2.3</emailID>
<participant>ID2</participant>
<sex>female</sex>
<age>16</age>
<education>12</education>
<L1>IrishEnglish</L1>
<date>2011-03-12</date>
<topic>MP</topic>
<salutation>Dear Mr Osbourne</salutation>
<body>I am writing etc.</body>
<closing></closing>
<time>3m00s/time>
</email>

This way, all participant data as well as their production data (+ some meta-data) are simultaneously available. (I noticed that some writers didn't set off the salutations and closings typographically, so whether these specific tags are useful is another matter.)

At the very least, though, I think it would be better if the participant data were saved in a spreadsheet rather than in a table in a docx file.

4. Data format: If you convert the table with the informant characteristics to a spreadsheet format (say CSV), please remove the last row with the summary statistics. (The reason is that if you read in a spreadsheet in a statistics program, it will assume that 'Mean (SD)' is the 43th participant, and that the column with Years of eduction is a string rather than a numerical variable.)

Minor comments - do as you see fit:

5. "used as a cut off" Perhaps 'baseline' rather than 'cut off'?

6. "Write an email to your MP". Perhaps clarify to non-British readers what MP stands for.

7. A minor pet peeve of mine: The informants' ages and years of eduction were obviously measured in years, but you report the means and standard deviations up to two decimal places throughout the manuscript. I realise this is common practice, but it's a case of false precision (http://en.wikipedia.org/wiki/False_precision). Would you consider rounding them off to the nearest integer (i.e. 14 +/- 3 rather than 13.36 +/- 3.30)? Similarly, 21.4% doesn't really carry any more information than 21% does.

Reviewer B: Kevin van Kalkeren
Review Completed: N/A

In their 2015 manuscript Thiel, Sage & Conroy provide emails written by 42 neurotypical participants. Each participant wrote three types of emails, which are included in the data set, as well as the task and demographic variables of the participants.

The data can be used in studies to find patterns in writing abilities. The paper is well-structured and concise, and very comprehensible. The only recommendation I would make, is leaving out the table with demographic variables, since it is already included in the data set itself, and means are reported in the text.

Even despite this, I would recommend publishing this paper, because I feel it could provide a convenient base for linguistic research and an insight in digital communication.