CSAFE Handwriting Database

Amy Crawford[1,2], Anyesha Ray[1], Alicia Carriquiry[1], James Kruse[1], Marc Peterson[1]

[1] *Iowa State University, Ames IA.*
[2] *Corresponding author, please email amy.m.crawf@gmail.com.*

Abstract

These data were collected to support the development of statistical approaches to the evaluation of handwriting as forensic evidence. Each enrolled participant provided handwriting samples at three data collection sessions, each at least three weeks apart. At each session participants completed a short survey and transcribed the contents of three prompts, each three times. This repository includes 27 scanned writing samples from each of 90 participants, making 2430 handwriting samples in total. In addition, survey data are available in table format including a few demographic variables and session specific information for each participant.

Access and Funding

A search-able version of this database will be made available at forensicstats.org/data.

This work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement #70NANB15H176 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, University of California Irvine, and University of Virginia.

Data Collection Methods

At each of three data collection sessions (**s01, s02, and s03**), participants completed a short survey and provided nine writing samples by transcribing three prompts, three times each. The first session was facilitated by a researcher to instruct and give time to answer questions regarding the informed consent process. Writing prompts are listed below with abbreviations used in the file naming structure.

LND: *The London Letter*[1], a common handwriting exemplar.

“Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zermott Street and then goes to Turin and Rome and will join Colonel Parry and arrive at Athens, Greece, November 27 or December 2. Letters there should be addressed King James Blvd. 3580. We expect Charles E. Fuller Tuesday. Dr. L. McQuaid and Robert Unger, Esq., left on the ’Y. X.’ Express tonight.”

WOZ: An excerpt from *The Wonderful Wizard of Oz*, by L. Frank Baum[2].

“Within a short time she was walking briskly toward the Emerald City, her silver shoes tinkling merrily on the hard, yellow roadbed. The sun shone bright and the birds sang sweet and Dorothy did not feel nearly as bad as you might think a little girl would who had been suddenly whisked away from her own country and set down in the midst of a strange land.”

PHR: A short common phrase.

“The early bird may get the worm, but the second mouse gets the cheese.”

During a session, the prompts were not written three times in succession, rather all three prompts were written one time for repetition #1 (**r01**), then all three prompts were written again for repetition #2 (**r02**), and a third time for repetition #3 (**r03**). The order in which participants were asked to transcribe the prompts within a repetition was block randomized according to their unique **writer identification number (WID)**, assigned to each writer upon enrollment.

Prior to each session a data collection packet was prepared for each writer. All necessary materials to complete the session were included, and the packets were mailed or delivered in-person to participants. The pages provided for completing writing samples were systematically generated using RMarkdown[3]. These pages were labeled with the intended writer identification number, prompt, and repetition number. This information was printed in plain text on the left side of a header and was embedded as a text string in a QR code to facilitate file naming on the right side.

Packets were mailed back to the researchers and the documents inside were scanned. A Python library[4] was used to extract embedded information from the QR code and an R[5] script used the text string to automatically assign a unique, informative name to each scanned document. The header was cropped off of each image using ImageMagick[6] functions to arrive at

[1] A.S. Osborn, 1929. Questioned documents, 2nd edn. New York, NY: Boyd Printing Co.

[2] L. F. Baum, 1900. The Wonderful Wizard of Oz, illustrated by W.W. Denslow. Chicago and New York: G.M. Hill Co.

[3] J. Allaire, Y. Xie, J. McPherson, J. Luraschi, K. Ushey, A. Atkins, H. Wickham, J. Cheng, W. Chang, R. Iannone, 2019. rmarkdown: Dynamic Documents for R. R package version 1.16, `https://github.com/rstudio/rmarkdown`.

[4] L. Hudson, 2019. pyzbar: Read one-dimensional barcodes and QR codes from Python 2 and 3. Python library version 0.1.8, `https://github.com/NaturalHistoryMuseum/pyzbar/`.

[5] R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

[6] ImageMagick Studio LLC, 2019. ImageMagick - Convert, Edit, or Compose Bitmap Images. ImageMagick release version 7.0.9-1. `https://imagemagick.org`.

the desired raw data images of the handwriting. Survey data entry was done manually, and was facilitated by an R Shiny[7] application.

File Formats and Naming

Writing Samples

Handwriting sample pages were scanned at 300dpi using an Epson DS-6500 document scanner and stored as images with the naming format of wAAAA_sBB_pCCC_rD.png, where,

- AAAA is a four digit writer identification number (WID) (between 0001 and 0180, not necessarily consecutive),
- BB is a two digit session number (01, 02, or 03),
- CCC is the three letter prompt shorthand (LND, WOZ, or PHR, see above), and
- DD is a number (01, 02, or 03), representing the repetition of a particular prompt in a given session.

Samples with writing very near the edge of the paper may not have been fully captured by the scanner. This occurred most often in the right or bottom margins. An extreme example of this occurs in file w0125_s01_pLND_r03.png. Images are stored in one of three folders based on their session number:

- **session1** is a folder containing files
  - w0001_s01_pLND_r01.png
  - w0001_s01_pLND_r02.png
  - w0001_s01_pLND_r03.png
  - w0001_s01_pPHR_r01.png

    ⋮

- **session2** is a folder containing files
  - w0001_s02_pLND_r01.png
  - w0001_s02_pLND_r02.png
  - w0001_s02_pLND_r03.png
  - w0001_s02_pPHR_r01.png

    ⋮

- **session3** is a folder containing files
  - w0001_s03_pLND_r01.png
  - w0001_s03_pLND_r02.png
  - w0001_s03_pLND_r03.png
  - w0001_s03_pPHR_r01.png

    ⋮

[7] W. Chang, J. Cheng, J. Allaire, Y. Xie, J. McPherson, 2019. shiny: Web Application Framework for R. R package version 1.40, `http://shiny.rstudio.com`.

Surveys

Survey data is stored in a table where there is one row for each participant. The file name is surveydata.csv. See the data dictionary:

Variable	Description
`wid`	Writer identification number, as described above.
`agegroup`	One of four age groups: `18-24`, `25-40`, `41-60`, `61+`.
`gender`	Gender identity. One of: `male`, `female`.
`hand`	Dominant writing hand. One of: `right`, `left`, `ambedextrous`.
`thirdgrade_usa`	`TRUE` if participant completed their third grade education in the United States, `FALSE` otherwise.
`thirdgrade_usa_region`	If `thirdgrade_usa` is `TRUE`, the region defined by the United States Census Bureau[8] (one of: `west`, `midwest`, `south`, `northeast`). If `thirdgrade_usa` is `FALSE`, then `NA`.
`s01_dae`	Number of days after enrollment that session #1 was completed. This column is all zero because enrollment occurs upon informed consent at the first session.
`s01_time`	Time of day session #1 was completed. See list of data value options below.
`s02_dae`	Number of days after enrollment that session #2 was completed.
`s02_time`	Time of day session #2 was completed. See list of data value options below.
`s03_dae`	Number of days after enrollment that session #3 was completed.
`s03_time`	Time of day session #3 was completed. See list of data value options below.

Any question left unanswered was coded as `NA`. Available values listed in the table above reflect the values that appear in the data, and do not necessarily represent the options listed on the surveys. The third grade location variables are included because children generally learn letter forms during early elementary education.

Data value options for the `sXX_time` variables are,

- `earlymorning`: earlier than 9:30am,
- `latemorning`: 9:30am – 12:00pm,
- `earlyafternoon`: 12:00pm – 2:30pm,
- `lateafternoon`: 2:30pm – 5:00pm,
- `earlyevening`: 5:00pm – 7:30pm, and
- `lateevening`: later than 7:30pm.

[8] United States Census Bureau. www2.census.gov.