Data Article

Dataset of coded handwriting features for use in statistical modelling

Anna Agius [a], Marie Morelato [a], Sébastien Moret [a],
Scott Chadwick [a], Kylie Jones [b], Rochelle Epple [b],
James Brown [c], Claude Roux [a],*

[a] University of Technology Sydney, Centre for Forensic Science, PO Box 123, Broadway 2007, Australia
[b] Forensics, Specialist Operations, Australian Federal Police, G.P.O. Box 401, Canberra, ACT, Australia
[c] University of Technology Sydney, School of Mathematical & Physical Sciences, PO Box 123, Broadway 2007, Australia

A R T I C L E   I N F O

A B S T R A C T

The data presented here is related to the article titled, "Using handwriting to infer a writer's country of origin for forensic intelligence purposes" (Agius et al., 2017) [1]. This article reports original writer, spatial and construction characteristic data for thirty-seven English Australian[1] writers and thirty-seven Vietnamese writers. All of these characteristics were coded and recorded in Microsoft Excel 2013 (version 15.31). The construction characteristics coded were only extracted from seven characters, which were: 'g', 'h', 'th', 'M', '0', '7' and '9'. The coded format of the writer, spatial and construction characteristics is made available in this Data in Brief in order to allow others to perform statistical analyses and modelling to investigate whether there is a relationship between the handwriting features and the nationality of the writer, and whether the two nationalities can be differentiated. Furthermore, to employ mathematical techniques that are capable of characterising the extracted features from each participant.

Specifications Table

Subject area	Forensic science
More specific subject area	Handwriting examination, forensic intelligence
Type of data	Table
How data was acquired	Manual coding
Data format	Categorical, numerical data
Experimental factors	Handwriting specimens were scanned and coded
Experimental features	Features of each participants' handwriting were selected, extracted and coded
Data source location	Handwriting specimens were collected from people living in the Sydney region (and whom had learnt to write English in New South Wales). All original documents are stored at the University of Technology Sydney (UTS), 15 Broadway, Ultimo, Sydney, NSW, 2007.
Data accessibility	The data is made available with this article
Related research article	Agius et al. [1].

Value of the data

- This data represents a complementary method for using handwriting features to obtain useful knowledge about the source.
- The data may help to expand the contributions of handwriting examination beyond answering the traditional Court-oriented questions and become more involved in a forensic intelligence framework.
- Researchers may use this data as a building point to trial different statistical techniques and modelling tools in order to determine whether any relationship exists within the dataset.

1. Data

The data is presented in Excel spreadsheets, where each column is headed by a particular handwriting characteristic with a description of the relevant codes for each feature. Table 1 below provides a description of the data in each spreadsheet. The first spreadsheet is titled "Writer & Spatial Characteristics". Writer characteristics are the traits specific to the individual writer and were taken from a survey filled out by each participant (see below). Spatial characteristics are the height and/or width relationships within and between individual characters and words, and how these are combined to form lines of words and paragraphs. The following six spreadsheets ('h' and 'th' are combined onto one spreadsheet) are the construction characteristics extracted for the characters 'g', 'h', 'th', 'M', '0', '7' and '9'. Construction characteristics are handwriting features that indicate how the writer forms a letter, number or symbol, e.g. number, position, order and direction of strokes. Please note, Vietnamese writer 11 was not included in the spreadsheets for the construction characteristics of the letters and numbers, as these characteristics could not be extracted from their handwriting due to the writing instrument that they used.

2. Experimental design, materials, and methods

The method for handwriting collection is conveyed in Agius et al. [1]. Briefly, seventy-four writers completed the handwriting collection package (37 Vietnamese and 37 English Australians). The

Table 1

Summary of the features coded in each of the Excel spreadsheets provided.

Spreadsheet title	Features coded
Writer & Spatial Characteristics	Writer's ID number; Nationality; Text indented, Valediction indented; Salutation indented; Paragraph 1; Paragraph 2; Slant; Handedness; Number of letters overhanging the right margin; Age; Years they have learnt English; Height of text; Width of text; Area of text; Average height of the letters 'a', 'e' and 'o'; and average number of letters per line
Lowercase 'g'	Writer ID number; Nationality; Handedness; Slant; Allograph; Letter number and the word the letter came from; Letter position within the word; Descending stroke design; Upper loop design; Connected to previous letter; Connected to subsequent letter; Direction of upper loop; Age; and the years they have learnt English
Lowercase 'h' and 'th'	Writer ID number; Nationality; Handedness; Slant; Allograph; Letter number and the word the letter came from; Letter position within word; Connected to previous letter; Connected to subsequent letter; Curve at the top of staff; Relative height of arch to staff; Arch retrace of staff; Arch shape; Height of 'h' staff relative to 't' staff; Age; Number of strokes to form 'th' combination; and the years they have learnt English
Uppercase 'M'	Writer ID number; Nationality; Handedness; Slant; Allograph; Letter number and the word the letter came from; Retrace on initial stroke; Apex 1 shape; Apex 2 shape; Apex 3 shape; Relative height of apex 2 to apex 1; Position of apex 2 relative to 1 and 3; Age; Number of strokes; and years they have learnt English
Number '0'	Writer ID number; Nationality; Handedness; Slant; Allograph; Number allocated to each '0' written by the participant; Start position; Finished position; Connection to start stroke; Embellishment; Direction; Age; and years they have learnt English
Number '7'	Writer ID number; Nationality; Handedness; Slant; Allograph; Number allocated to each '7' written by the participant; Style; Curve at bottom of staff; Angle between the horizontal and vertical strokes; Age; Number of strokes; and years they have learnt English
Number '9'	Writer ID number; Nationality; Handedness; Slant; Allograph; number allocated to each '9' written by the participant; Upper loop design; Descending stroke design; Direction of upper loop; Age; Number of strokes; and years they have learnt English

Fig. 1. Source document (left) and provided lined piece of white A4 copy paper to write out the source document (right).

UNIVERSITY OF
TECHNOLOGY SYDNEY

Handwriting Specimen Form 2A

_____/_____/_____/_____/_____/
A   B   C   D   E

_____/_____/_____/_____/_____/
F   G   H   I   J

_____/_____/_____/_____/_____/
K   L   M   N   O

_____/_____/_____/_____/_____/
P   Q   R   S   T

_____/_____/_____/_____/_____/
U   V   W   X   Y

_____/_____/_____/_____/_____/
Z   a   b   c   d

_____/_____/_____/_____/_____/
e   f   g   h   i

_____/_____/_____/_____/_____/
j   k   l   m   n

_____/_____/_____/_____/_____/
o   p   q   r   s

_____/_____/_____/_____/_____/
t   u   v   w   x

_____/_____/
y   z

_____/_____/_____/_____/_____/
0   1   2   3   4

_____/_____/_____/_____/_____/
5   6   7   8   9

_____/_____/_____/_____/_____/
10   00   11   22   33

_____/_____/_____/_____/_____/
44   55   66   77   88

_____/_____/
99   100

_____
276 Punchbowl Road, Punchbowl, NSW

_____
95/59 Unit Broadway Avenue, Sydney NSW

_____
1373 Jones Street Kingsgrove, NSW

_____
45/67 Unit King George Road, Sydney, NSW

_____
PO Box 458 Campsie Avenue, NSW

_____
1009 Campbell Street, Campelltown, NSW

Fig. 2. Upper and lowercase letters, numbers and postal addresses written out by participants.

participants were required to copy out the source document (Fig. 1); the alphabet in upper and lower case and the numbers zero to ten (Fig. 2); and postal addresses (Figs. 2 and 3). The source document contained all twenty-six letters of the alphabet, upper and lowercase, and the numbers 0–9. It was copied down onto a 24-lined piece of A4 white copy paper with the lines 0.8 cm apart and a 2 cm margin [1].

UNIVERSITY OF TECHNOLOGY SYDNEY

Handwriting Specimen Form 3A

FROM: (please PRINT) PHONE	TO: (please PRINT) PHONE
Mr. Bac Luan Nguyen 700 Tran Hung Dao, Ward 2, District 5, Ho Chi Minh city, Vietnam.	Mrs. Phien Phien Tran 123 Peptone Crescent, Cabramatta West, 2166, NSW, Australia.
FROM: (please PRINT) PHONE	TO: (please PRINT) PHONE

Fig. 3. Space provided for participants to write out two postal addresses.

Salutation Indent

Paragraph 1

Text body Indent

Paragraph 2

Valediction Indent

Letters overhanging the right margin (if any)

Height of text

Width of text

Dear Zach,

Well, the old class of "16" is through at last. You ask where the boys are to be. Val Brown goes on the 24th to Harvard for law. Don't forget to address him as "Esquire." Ted Updyke takes a position with the N. Y. W. H. & H. R. R., 892 Ladd Ave., Fall River, Massachusetts, and Jack McQuade with the D. L. & W. at Jersey City, N. J. 400 E. 6th Street. William Fellows just left for a department position in Washington; his address is 735 South G. St. At last account, Dr. Max King was to go to John Hopkins for a Ph.D. degree. Think of that! Elliott goes to Xenia, Ohio, to be a Y. M. C. A. secretary. I stay here for the present. What do you do next? How about Idaho?

Yours truly, and goodbye.

Fig. 4. Diagram of the spatial features recorded from the source document.

The handwriting features that were chosen for extraction were separated into writer, spatial and construction characteristics. They were recorded or coded depending on the type of data, i.e. the age of the writer was recorded as is, however descriptive information like the direction of the stroke was coded zero for anticlockwise and one for clockwise. The spatial characteristics were extracted only from the source document as it had a sufficient quantity of writing as opposed to the other pages of the collection package. Fig. 4 is a diagram demonstrating the spatial characteristics extraction process.

Table 2
Qualitative spatial characteristic codes.

Spatial characteristic	Code	Code description
Salutation indented	0	No indent
	1	Indent
Text indented	0	No indent
	1	Indent
Valediction indented	0	No indent
	1	Indent
Paragraph (1)	0	No space between salutation and text body
	1	Space between salutation and text body
Paragraph (2)	0	No space between text and valediction
	1	Space between text body and valediction
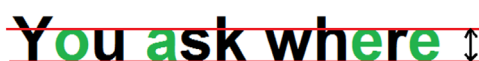
Fig. 5. Example of measuring the average height of 'a', 'e' and 'o' from the text. The distance between the top and bottom lines of best fit (red lines) was measured.

Table 3
The number of characters that had their construction characteristics extracted and coded.

| Character	Count
g	10
h	28
th	13
M	8
0	12
7	8
9	7

Table 2 summarises the codes of the qualitative spatial characteristics. Fig. 5 provides a visual depiction of how the feature, the average height of 'a', 'e' and 'o', was measured.

The construction characteristics for 'g', 'h' and 'th' were extracted from the source document and from the individual upper and lowercase sections in Fig. 2. They were not extracted from the postal addresses in Fig. 2 as the writing tended to be cramped because there was not much space to write out the postal address, or Fig. 3 as many of the participants wrote the addresses in uppercase. The characters 'M', '0', '7' and '9' were taken from all three forms in the collection package. Table 3 shows how many of each character had their construction characteristics coded in the collection package for each person.

Examples of the construction characteristics extracted for each character and their numerical code are presented in Table 4. This is provided in order to visually convey how the data was created, allowing researchers to make an informed selection of (a) statistical technique(s) which may be more suitable to analyse and model the data.

Table 4
Examples of all of the construction characteristics extracted for each character and their respective codes.
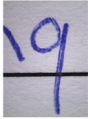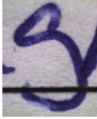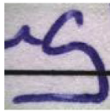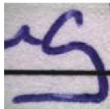
Character	Construction Characteristic	Code	Example
'g'	Descending stroke design	0 – Straight	
		1 – Open	
		2 - Closed	
	Upper loop design	0 – Open	
		1 - Closed	
	Connected to previous letter	0 – Not connected	
		1 - Connected	
	Connected to subsequent letter	0 – Not connected	
		1 - Connected	

Table 4 (*continued*)
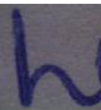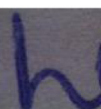
	Direction of upper loop	0 – Anticlockwise	
		1 - Clockwise	
'h'	Connected to previous letter	0 – Not connected	
		1 - Connected	
	Connected to subsequent letter	0 – Not connected	
		1 – Connected	
	Curve at top of staff	0 – Left	
		1 – Right	
		2 – Loop	
		3 – Straight	

Table 4 (*continued*)
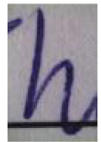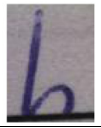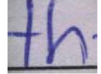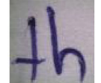
	Relative height of arch to staff	0 – Lower half	
		1 – Half way	
		2 – Upper half	
	Arch retrace of staff	0 – Closed retrace	
		1 – Open retrace	
		2 – Not connected	
	Arch shape	0 – Rounded	
		1 - Angular	
'th'	Height of 'h' staff relative to 't' staff	0 – 't' above 'h'	
		1 – Equal	
		2 – 'h' above 't'	

Table 4 (*continued*)
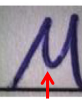
	Number of strokes	1	
		2	
		3	
'M'	Retrace on initial stroke	0 – Closed retrace	
		1 – Open retrace	
		2 – No retrace	
	Number of strokes	1	
		2	
	Apex 1 shape	0 – Rounded	
		1 - Angular	
	Apex 2 shape	0 – Rounded	
		1 - Angular	

Table 4 (*continued*)
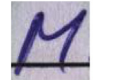
	Apex 3 shape	0 – Rounded	
		1 - Angular	
	Relative height of apex 2 to apex 1	0 – Above half way	
		1 – Half way	
		2 – Below half way	
	Symmetry of apex 2 relative to 1 and 3	0 – Left	
		1 – Middle	
		2 - Right	
'0'	Start position	0 – Upper left	
		1 – Middle top	
		2 – Upper right	
		3 – Bottom left	

Table 4 (*continued*)
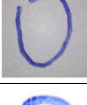
		4 – Middle bottom	
		5 – Bottom right	
		6 – Middle of circle	
	Finished position	0 – Overlap	
		1 – No overlap	
	Connection to start stroke	0 – Connected	
		1 – Not connected	
	Embellishment	0 – No embellishment	
		1 – Slash	
		2 - Dot	
	Direction of stroke	0 – Anticlockwise	
		1 - Clockwise	

Table 4 (*continued*)
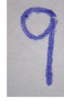
'7'	Style	0 - Classic	
		1 – Crossbar	
		2 – Serif only	
		3 – Serif and crossbar	
		4 – Reverse serif	
		5 – Crossbar and reverse serif	
	Number of strokes	1	
		2	
	Curve at bottom of staff stroke	0 – Right	
		1 – Straight	
		2 - Left	
	Angle between horizontal and vertical strokes	0 – Rounded	
		1 – Angular	

Table 4 (*continued*)

		2 - Loop	
'9'	Upper loop design	0 – Open	
		1 – Closed	
	Descending stroke design	0 - Left	
		1 – Straight	
		2 - Right	
	Direction of upper loop	0 - Anticlockwise	
		1 - Clockwise	
	Number of strokes	1	
		2	

Transparency document. Supplementary material

Transparency document associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.12.014.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.12.014.

Reference

[1] A. Agius, M. Morelato, S. Moret, S. Chadwick, R. Epple, K. Jones, J. Brown, C. Roux, Using handwriting to infer a writer's country of origin for forensic intelligence purposes, Forensic Sci. Int. 282 (2018) 144–156.