iUF Firemaker: A benchmark data set for writer identification

Lambert Schomaker & Louis Vuurpijl

November 2000

Abstract

This report is the result of a collaboration between the Forensic Institute (NFI) of the Dutch Ministry of Justice and the NICI, Nijmegen University, The Netherlands. The goals of this project were twofold: (a) The collection of an independent high-quality test set of handwritten data for the purpose of comparing available systems for forensic writer identification, and (b) the comparison of two such systems. In this report a description of the collection process and the collected database is given. It is intended to serve as documentation along with the image database.

Acknowledgements

The authors would like to thank the following persons for their support: Martijn Huygevoort (for the collection process), Merijn van Erp and Fusi Wang (for data labeling and measuring), Hartmut Giessen, Katrin Franke (for many stimulating discussions on forensic handwriting research), Jurrien Bijhold, and his ideas on how to compose the textual material for the test), Wim de Jong, the printers at Ipskamp (who most probably never had a stranger customer than us), the persons providing their handwritten samples, Charles de Weert (for his knowledge on optics and color), Carol Clements at Fujitsu (for providing information on dropout colors and the spectral characteristics of the scanner lamp), Janek Mackowiak and Jeroen Beekhuis for computer system maintenance, the students Arie Baris and Maartje van Hardeveld and many others who assisted alongside the road during this difficult journey.

Contents

Chapter 1

Introduction

This report is the result of a collaboration between the Netherlands Forensic Institute (**NFI**) of the Dutch Ministry of Justice and the Cognitive Engineering Group of the Nijmegen Institute for Cognition and Information of Nijmegen University, The Netherlands (**NICI**). The goals of this project were twofold: (a) The collection of an independent high-quality test set of handwritten data for the purpose of comparing available systems for forensic writer identification, and (b) the comparison of two such systems. In this report, only a description of the collection process and the collected database is given.

At the end of the project with the **NFI**, it was decided to make the scanned image data available to the scientific community at large. The database was dubbed 'Firemaker', a contraction (in English) of the names of the principal investigators. The distribution will be realized by the International Unipen Foundation (iUF).

Chapter 2

Data collection

In this chapter, the details of the data collection process and of the collected handwritten samples are described.

2.1 Requirements

In early discussions with project partner **NFI**, a number of requirements for the handwritten samples were determined:

- high-quality scans of sufficient resolution
- formats and resolution suitable for target writer identification systems
- reusability for the coming years
- comparability with earlier collections, like those routinely recorded by police departments in the Haaglanden region.
- a sufficient number of writers (in the hundreds)
- text content with rich variation in letters, digits, upper and lower case
- conditions of normal and forged (i.e., disguised) handwritten style
- conditions of text copying and self-generated text

Other requirements involved the presence of lineation guidelines on the response sheets. Although vertical line distance in self-generated handwritten text may be a discriminatory writer characteristic, it was deemed more practical to use lineation guidelines on the response sheet. However, together with the other information on the response sheet, this could result in a low quality of the scans, with pixels related to the ball-point ink being mixed with pixels from the lineation, boxes and other content. It was decided to solve this problem using a suitable dropout color. A dropout color is a color which is almost fully reflected by the light spectrum emitted by the scanner lamp such that it has the same sensed luminance as the white background under ideal conditions. For automated processing and later reuse of the data, it was also necessary to add essential information in a systematic way on each response sheet. This information consists of writer age and gender, handedness, a simple machine-readable code identifying the text content, and a unique number for each writer (i.e., individual data set).

2.2 Text content and response sheet layout

The decision was made to require the subjects to write on four different A4 pages. Instead of contiguous lines of text in a full paragraph, the texts where organized as blocks as much as possible, to facilitate (semi)automatic processing at a later stage. Instructions to the writer (in Dutch) are given in Appendix I. Writers were specifically instructed to use their own word layout, at the same time avoiding hyphenations at the end of lines, such that full, single-block words are produced. The text content is adapted from the 'Haaglanden instructions' in order to contain all letters of the alphabet, and includes a new condition. The writing conditions where defined as follows:

2.2.1 Text1. Normal handwriting style

This condition is intended to elicit a normal handwriting style in a visual text-copying context: The text is presented in the form of machine-print characters (as opposed to, e.g., an auditory dictation task).

> Bob, David en sexy Xantippe sparen postzegels van de landen Egypte, Japan, Algerije, de USA, Holland, Italië, Griekenland en Canada.
>
> Zij bezochten veilingen en reisden met de KLM. Voor korte afstanden huurden ze een auto, meestal een VW of een Ford.
>
> De veilingen waren van 7-4-1993 tot 3-5-1993 in New York, Tokyo, Québec, Phoenix, Rome, Parijs, Zürich en Oslo.
>
> Omdat de veilingen steeds begonnen om 12 uur en je gemiddeld 200 tot 300 kilometer moest rijden, stonden zij steeds om 6.30 uur op en vertrokken om 8 uur uit het hotel.
>
> Elke dag hadden ze vijfhonderd (f 500,-) gulden nodig. Daarvoor gebruikten ze elke keer een cheque van tweehonderd (f 200,-) en een cheque van driehonderd (f 300,-) gulden. Aan geschenken gaven ze ongeveer honderd gulden (f 100,-) uit.

2.2.2 Text2. Upper-case (block print) style.

On this sheet, the subjects were asked to produce a block-print style, of which an example was given in the writer's manual (Appendix I.).

> NADAT ZE IN NEW YORK, TOKYO, QUÉBEC, PARIJS, ZÜRICH EN OSLO WAREN GEWEEST, VLOGEN ZE UIT DE USA TERUG MET VLUCHT KL 658 OM 12 UUR.
>
> ZE KWAMEN AAN IN DUBLIN OM 7 UUR EN IN AMSTERDAM OM 9.40 UUR 'S AVONDS. DE FIAT VAN BOB EN DE VW VAN DAVID STONDEN IN R3 VAN HET PARKEERTERREIN. HIERVOOR MOESTEN ZE HONDERD GULDEN (F 100,-) BETALEN.

2.2.3 Text3. Forged style

In this condition, the writers were asked to produce handwriting in a style "as if someone else had written the text".

> Nog dezelfde avond reden ze naar hun vrienden Chris, Emile, Jan, Irene en Henk, nadat ze hun vriendinnen Greta en Maria hadden opgehaald.
>
> Samen hadden ze vijfhonderd (500) zeldzame postzegels gekocht, Bob driehonderd (300) en David tweehonderd (200).
>
> De reis was de moeite waard geweest.

2.2.4 Text4. Self-generated, natural handwriting

In order to elicit natural, self-generated handwriting behavior, a condition was added in which writers were asked to describe the content of a given cartoon in their own words (see Appendix I.).

2.3 Physical conditions

The surroundings during collection were sufficiently lit, either by artificial light or by daylight coming from above. The support surface was a regular rigid table top. A thin light-blue cardboard sheet was

placed between table and response sheet to ensure standardized support conditions. This means, for instance, that ink trace thickness variations will be more due to writer differences than due to the recording conditions. The same type of pen was used by all subjects (Figure 2.1).

Figure 2.1: The Soennecken No. 1 M (Din 16554) ball-point pen

Coated A4 paper was used with low glare and optimal ink adhesion properties for the used type of pen. At pen-down there will be a short 'roll-in' trajectory which is typical for ball point pens. At pen-up, a similar 'roll-off' tapering will be present. By using grey-scale scanning these boundary phenomena will be retained in the data.

2.4 Writers

Writers were recruited from a student population, at the end of lectures in large lecture halls, or on the basis of individual appointments at a fixed location. Table 2.1 shows some statistics on the writers which took part in the experiment. Writers were predominantly (but not exclusively) right handed and younger than 30 years of age.

Table 2.1: Some statistics on the writers who took part in the data collection process.

Total number of writers	431	persons
Male	141	persons
Female	290	persons

Table 2.2 and 2.3 give more detailed information.

	male	**female**	**total**
Total number of writers	101	301	402
Average age	25	23	24
Right handed	92	265	357
Left handed	9	36	45

Table 2.2: Statistics for all 403 collected writers

	male	**female**	**total**
Total number of writers	82	168	250
Average age	25	23	24
Right handed	74	147	221
Left handed	8	21	29

Table 2.3: Statistics for the 250 scanned writers

2.5 Equipment

Figure 2.2 gives an overview over the equipment used in this project.

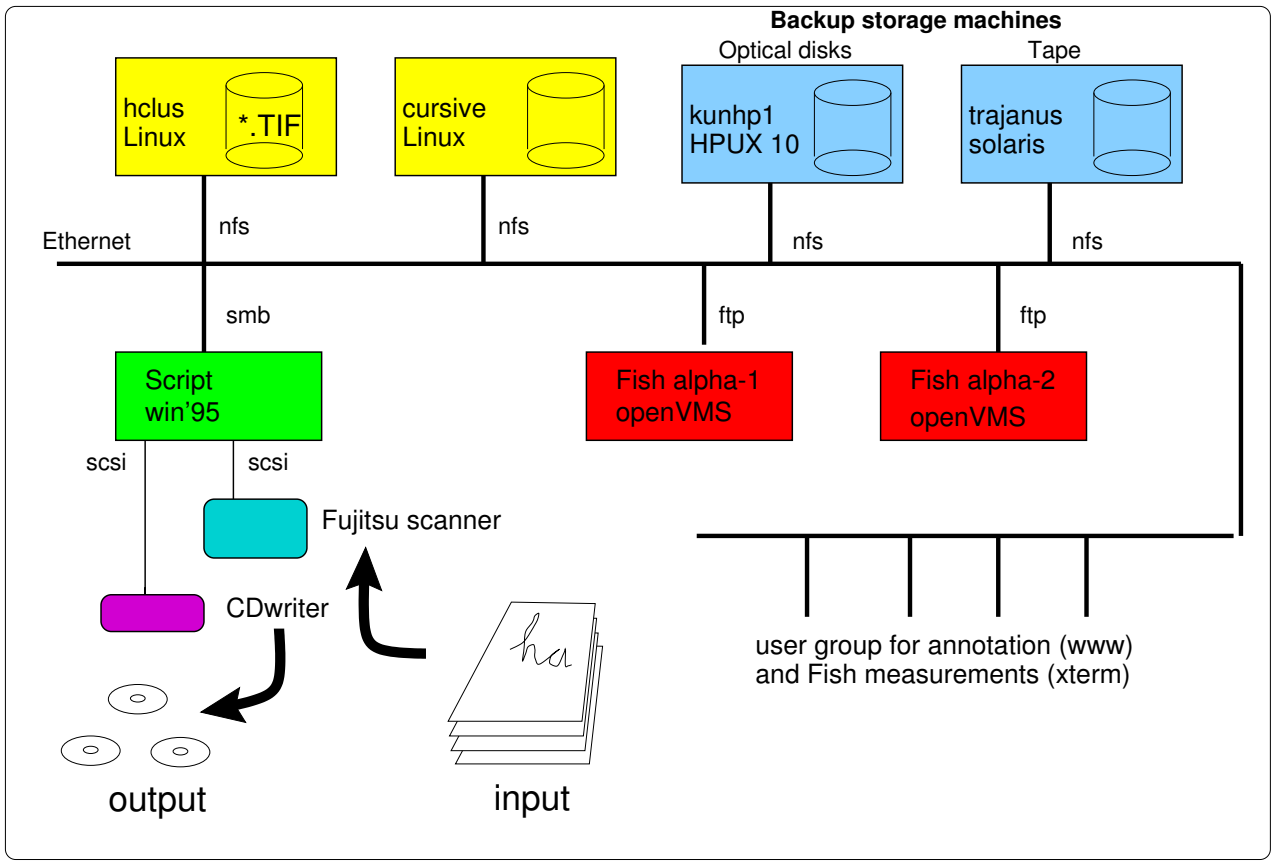


Figure 2.2: An overview of the system architecture

At the input side, middle left in Figure 2.2, there is a Fujitsu M3093DG scanner. The relevant technical specifications are given in Table 2.4. This industrial-quality scanner was chosen because of the combination of high scanning quality with robustness and scanning speed. This scanner is connected via a SCSI cable to a Compaq 75 MHz/81MB PC, dubbed Script. This system also has a CD-R writer connected to it. This system runs Microsoft Windows '95. The ScandAll V2.5 scanning software was used. The scans were copied to a Linux box (dubbed Cursive), using a mounted Network Neighbourhood disk made available through the SMB disk-access compatibility software on the Linux box.

Table 2.4: Fujitsu M3093DG scanner specifications

Sensor	Dual charge coupled device (CCD) image sensor
CCD Resolution	400 (dpi)
Interpolated resolutions	600, 300, 240, 200, 150, 100 (dpi)
Grayscale	8 bits at 400 dpi
Scanning Speed	27 ppm at 200 dpi (letter)
Document Size	A4
Interface	SCSI-2
Document feeding mode	Flatbed and automatic document feeder (ADF)
ADF capacity	50 sheets

2.6 Procedure

Scanning was done at 300dpi, 8-bit grey levels. Figure 2.3 shows an enlargement of a typical 300dpi scan. Note the absence of background noise pixels and the grey levels which are discernible within the ink trace.

Figure 2.3: The word *Bob*, extracted and enlarged from a typical 300dpi scan

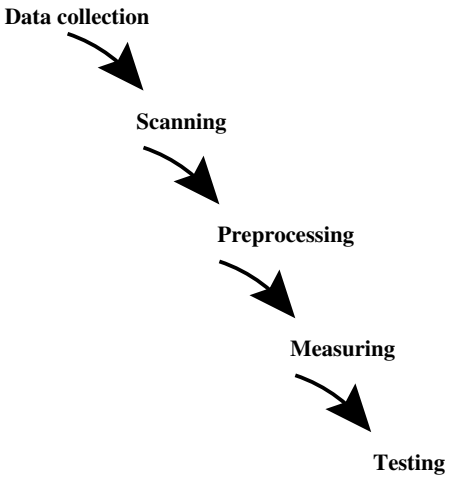


Figure 2.4: The processing pipeline

Figure 2.4 shows the global processing pipeline. A different view on the same process is given

in Figure 2.5. This figure shows that for a collection of writers, each writer produces four texts. A text is divided into blocks. Within each block, the actual measurements take place. The raw scans (*.tiff.gz) are stored permanently and are used for a number of exported selections. For instance, the extraction of selected text blocks before sending the image to a writer identification was very useful because it (a) reduced the amount of computation time considerably, and (b) because it allowed for a recombination of text blocks into larger text images. A GZIP-compressed .tiff scan contains 318kB on average, whereas the raw .tiff scans contain 8.3 MB of data.

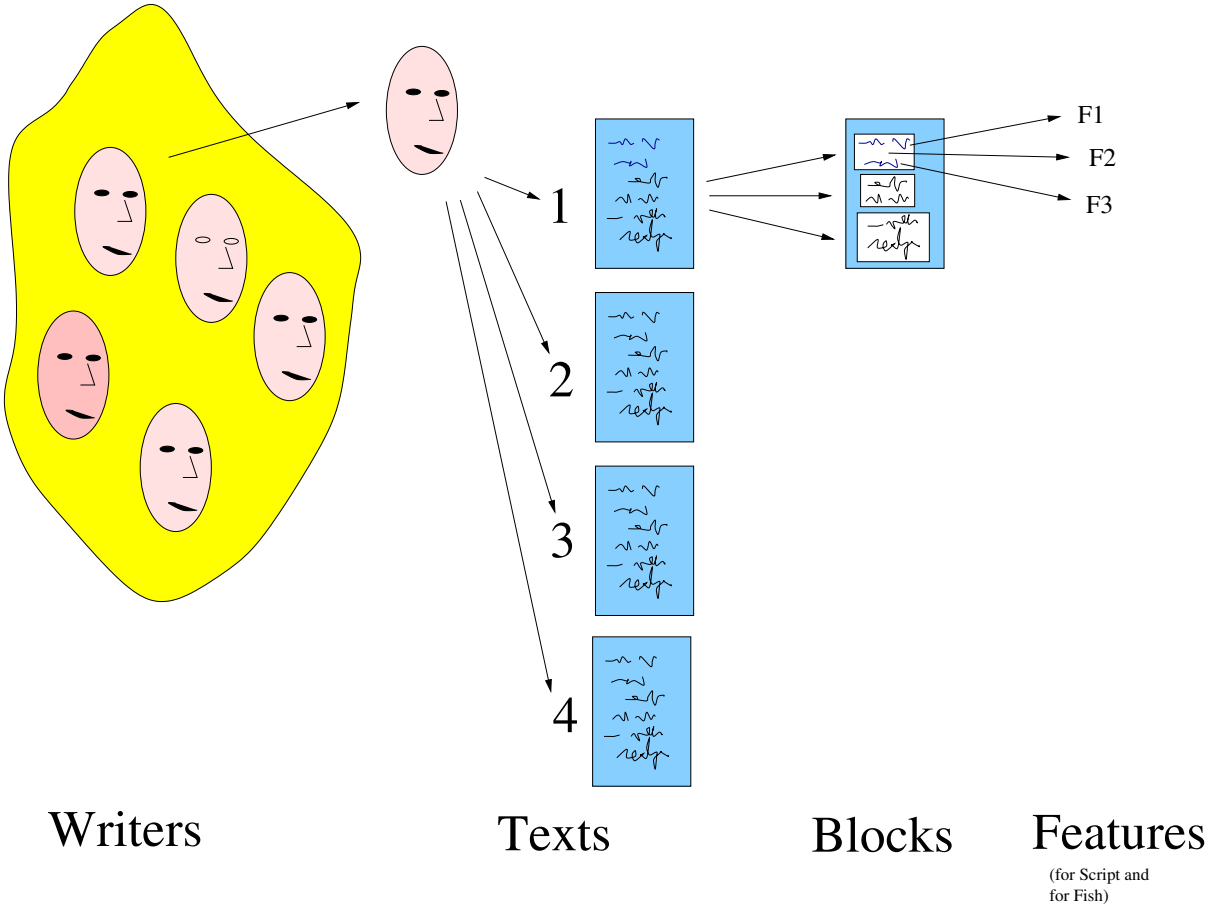


Figure 2.5: Another view on the processing pipeline

After scanning, each scan was checked automatically for a number of possible errors. Categories of interest are: Resolution, Gender, Form identification (i.e., Text1, Text2, Text3, Text4), and presence of ink where it was expected on the page. Although some experiments with despeckling and background removal were performed, the scans are of such a quality that these image operations were **not** necessary. The ASCII representation of Text1 through Text3 is more or less fixed. Figure 2.6 shows an example of a scan during the (fast) annotation of the text blocks by tagging each block with the given expected ASCII representation. A paragraph is selected with the mouse and tagged with the ASCII text as prompted in the writer's manual. Note that this process does not take into account possible spelling errors produced by the writers. At the bottom of the page, the text-code block ▪▫▫▫ is visible, as well as a 3cm calibration mark ⊢——⊣. The number (990302) at the bottom of the page identifies the set of four pages produced by this writer and is suitable for OCR processing. At the top of each page are the writer-identification data. The printed ink color of these fixed components on the form was

lighter than the ink of the pen in order to prevent suboptimal grey-level scaling for the ball-point ink by 'intelligent' scanner-driver software. Since the text of the 'cartoon condition' (Text4) is completely unknown, it had to be annotated manually. This was done using a group-collaboration system based on standard Internet browsers and a CGI interface (Figure 2.7). Finally, paragraphs of text were extracted and aligned at the writing line.

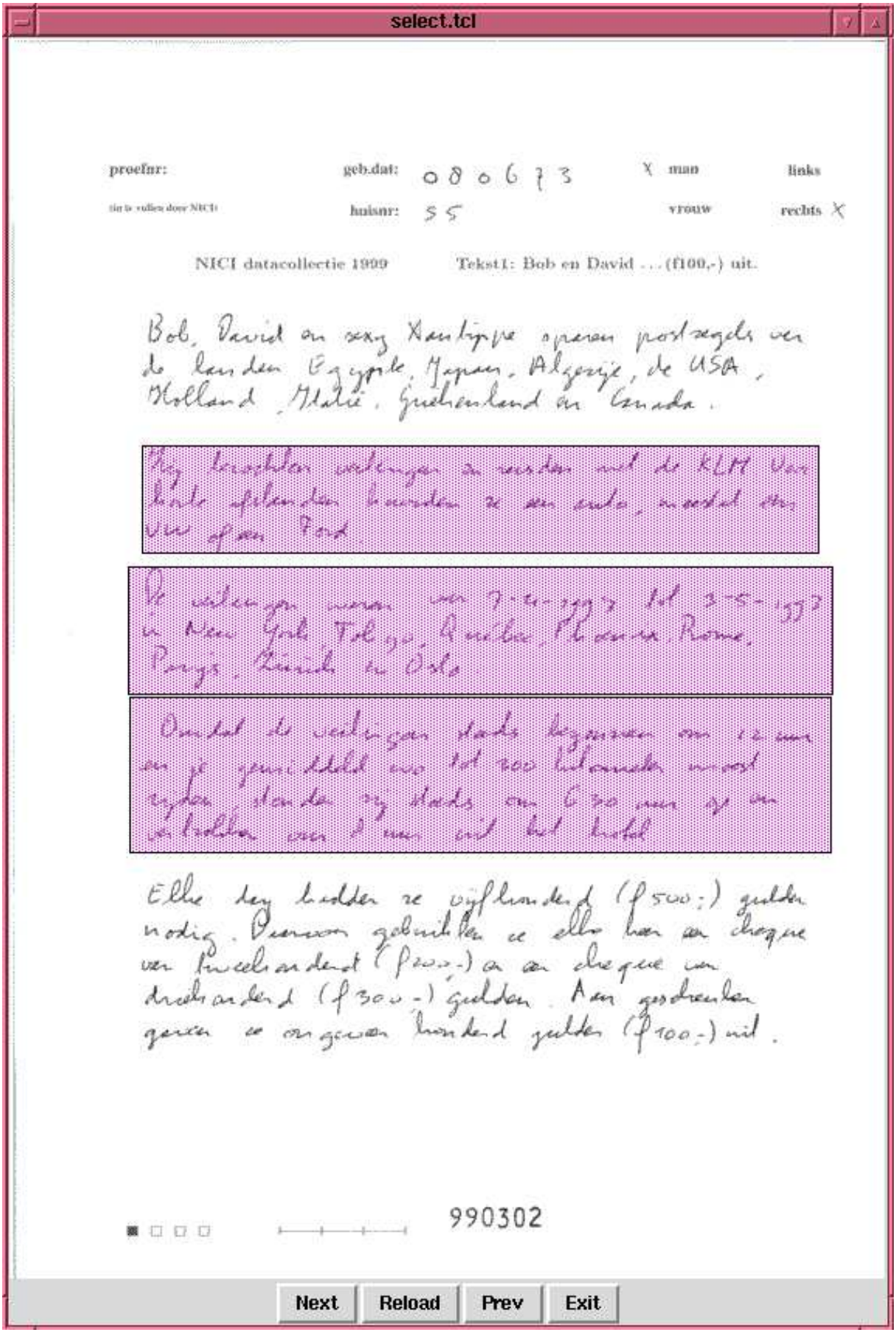


Figure 2.6: A screen dump of a *Text1* scan during quick annotation (see the text for more details).

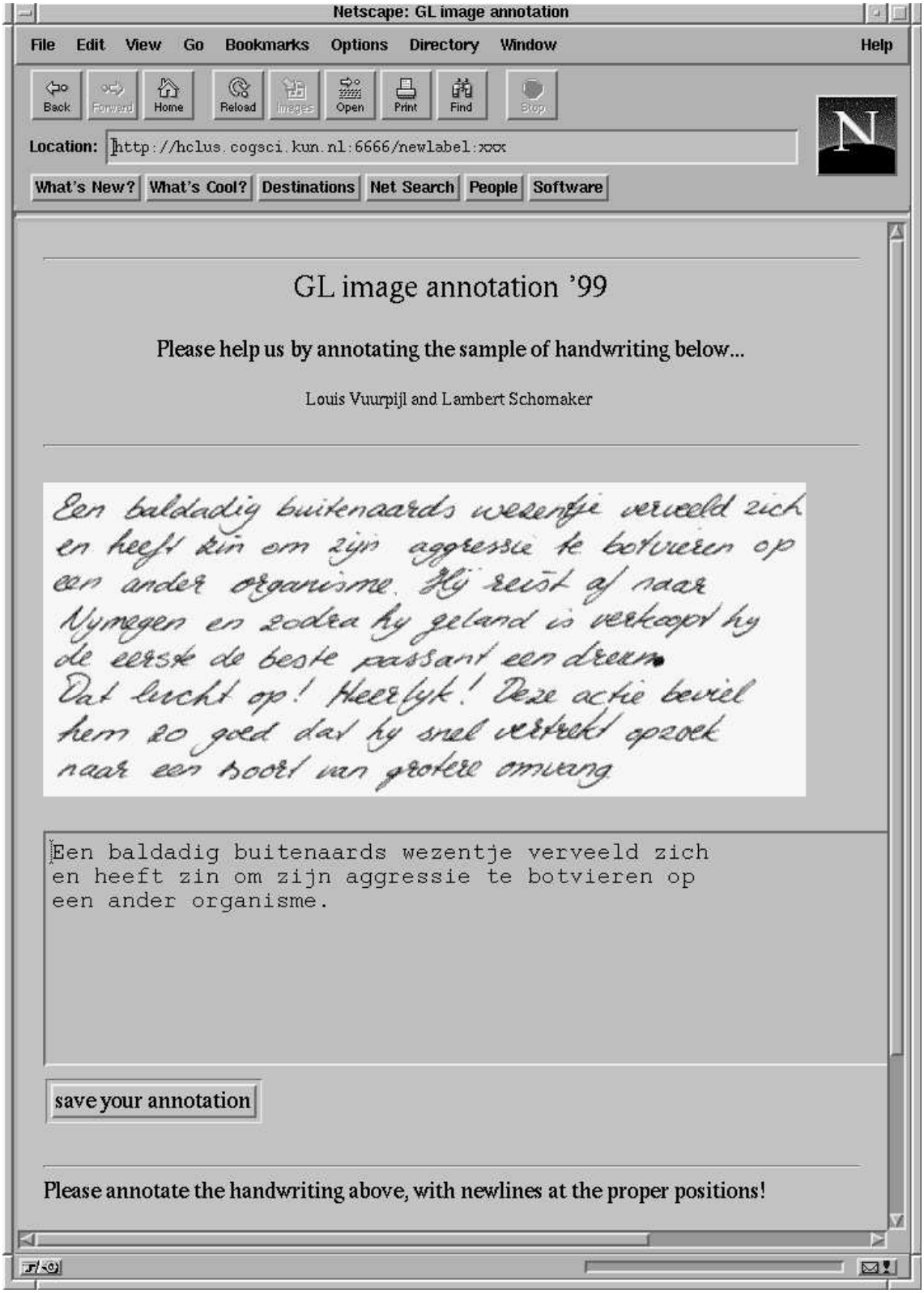


Figure 2.7: A screen dump of a manual *Text4* annotation by means of a WWW browser. This was done by prompting users in the working group via periodical emails containing the relevant URL to a CGI script and a kind request to participate in the annotation process.

2.7 Software

This section describes the available and developed software required for the two processing routines for the forensic writer identification as executed in the project with **NFI**. Scanning was done at 200dpi and 300dpi. Only the 300dpi data were entered into the iUF Firemaker distribution. Each image is stored in the `.tiff` image format. Existing software was used for visualization of the scanned images (`xv`[1]), and for loss-less conversion of one image format to the other (`ImageMagick`[2]). Furthermore, dedicated software was developed for *validating*, *annotating* and *processing* the scanned documents. We would like to emphasize here that the software described here was specifically developed for this project and every component was absolutely required to establish the goals of a high-quality comparison of writer identification systems.

2.7.1 Validation of scanned documents

Validation of the scans would check for the following requirements:

1. **Image size** For a A4 scan at 200dpi, the required image size must be 1653x2338 pixels and for a resolution of 300dpi, it is 2480x3507 pixels large.

2. **Image calibration** In some occasions, a scan can be slightly rotated or translated with respect to the scanning area. By considering the known stable points, such as the four rectangles identifying a document and the rectangles for marking gender and handedness of a writer, it can easily be verified if a scan is calibrated or not.

3. **Gender and handedness verification** The same software as used for validating the image calibration was used to determine if a writer has indicated his/her gender and handedness. Knowing where the rectangles for marking gender and handedness are, it can be determined which one is marked by just counting the black pixels in the area of interest.

4. **Verification of identification number** Each document is identified by a 6-digit number. In our databases, writers are identified based on the last 3 digits of the number. For example `152a` identifies the first document produced by writer `990152`, which is stored under the name of `152a.tif`. Software was produced to recognize the last three digits of the document number, and prompt for further investigation if the recognized number does not match the number under which it was stored. Matching the three digits was performed with a four-layer MLP neural network.

The produced software for validating scanned documents is completely written in the C programming language and tested on various Unix platforms.

`validate_calibration` Searches for rectangles related to gender, handedness and document identification number. If the image size is not as expected, an error is issued.

`extract_gender` Searches for the rectangles related to gender, and outputs either "F" or "M' or "U". In the latter case, a message is issued that no gender information can be extracted.

`extract_handedness` Searches for the rectangles related to handedness and outputs either "L" or "R' or "U". In the latter case, a message is issued that no handedness information can be extracted.

`verify_identification` Extracts the last three bitmaps of the document's identification number and recognizes each of these based on a pre-trained MLP. If the expected number and the recognized number does not match, an error is issued.

[1] `xv` is available at `http://www.trilon.com/xv/xv.html`

[2] `ImageMagick` is available at `http://www.wizards.dupont.com/cristy/ImageMagick.html`

2.7.2 Annotation of scanned documents

Figure 2.6 depicted in the previous section shows a tool for indicating the bounding box of paragraphs in a document. Although software was developed for automatically extracting paragraphs from a document, based on the horizontal frequency of "ink" on a line, it was decided that manually annotating the bounding box of each paragraph would result in even more reliable paragraph boundaries. For each scanned document, the bounding box of the known paragraphs were stored. This tool was written in C and Tcl/Tk.

A second tool is depicted in Figure 2.7. Using this tool, documents in the category "unconstrained handwriting" can be annotated with the correct text. For this setup, a dedicated web server was developed which depicts a sample of the (unknown) text produced and a window via which a user can input the recognized text.

`annotate_paragraphs` Reads a scanned document and displays it. Using the mouse, a user can select a bounding box and label it with the corresponding text.

`web_annotator` Is a dedicated web server written in C. The server waits for browser-requests for labeling an image. Upon this request, a <html> page is dynamically produced containing the image to be annotated and a <form> via which the annotated text can be submitted.

2.7.3 Processing of scanned documents

Several tools were written. One will be mentioned here:

1. **Paragraph extraction** For each scan, paragraphs were manually annotated using the software tool depicted in Figure 2.6. Software was developed to extract one or more paragraphs from a document, justified on the writing lines. A more elaborate description of this process is given elsewhere.

`extract_paragraphs` Reads a scanned document and bounding boxes of paragraphs to be extracted. Outputs a new image with paragraphs adjusted on writing-line boundaries.

2.7.4 Extraction of paragraphs

The four texts to be produced by writers was generated in close collaboration with the **NFI**. Based on a further discussion, it was decided that for the first series of tests a comparison would be made between two paragraphs from text1 as "setA" and the remaining three paragraphs of text1 as comparison "setB". For the two setA paragraphs, "Bob ... Canada" and "Elke ... uit" were chosen. The remaining three paragraphs "Zij ... Ford", "De ... Oslo" and "Omdat ... hotel" were selected to comprise the setB paragraphs.

Software was developed to extract a sub-image from a scanned document based on the specified paragraph bounding boxes and adjusted on writing line boundaries as described by the (known) document layout. As an example, consider Figure 2.6, which depicts the first text produced by writer 302. Using the program `extract_paragraphs`, the following image is produced:

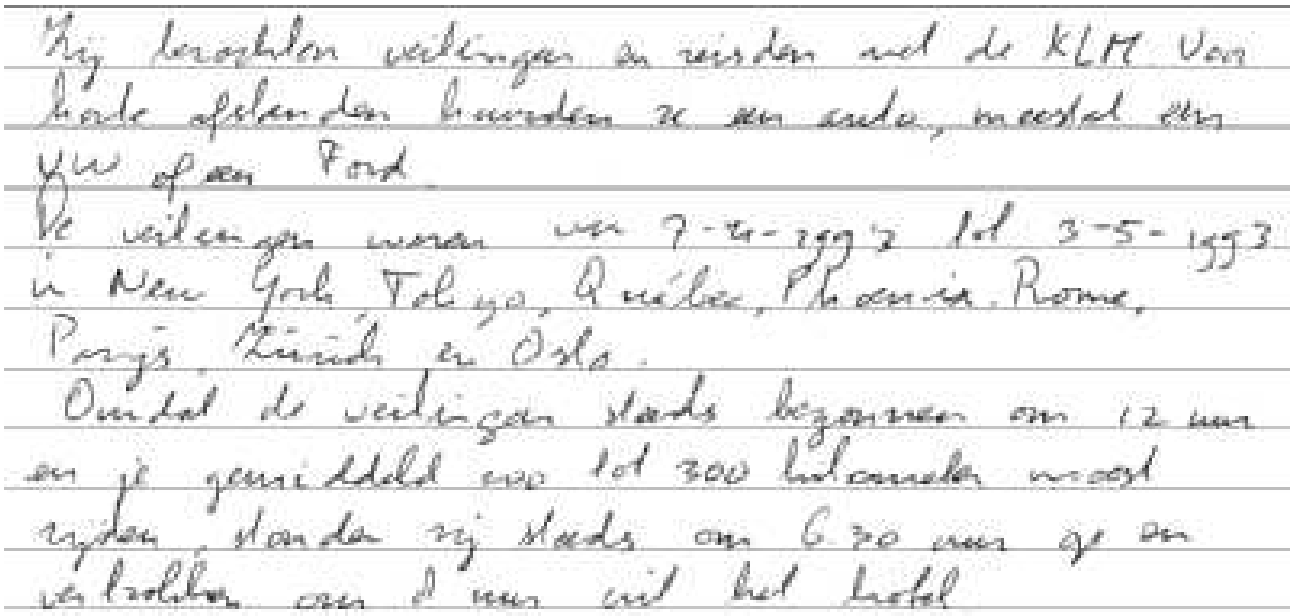


Figure 2.8: *Extracted paragraphs using* `extract_paragraphs`. *Note that in this image, the writing-lines are inserted, whereas for the actual images to be measured this is not performed.*

Chapter 3

CDROM Disk Structure

The iUF Firemaker database is available in the form of four separate sets. Each set corresponds to a page and writing condition. The directory structure is as follows:

```
300dpi/
  p1-copy-normal/       Copying task, normal writing style
    15201.tif             Writer 152 (page 1)
    15301.tif             Writer 153 (page 1)
      .
      .

  p2-copy-upper/*2.tif  Copying task, UPPER-case
    15202.tif             Writer 152 (page 2)
    15302.tif             Writer 153 (page 2)
      .
      .
  p3-copy-forged/*3.tif Copying task, instructed to mimic another script style
    15203.tif             Writer 152 (page 3)
    15303.tif             Writer 153 (page 3)
      .
      .

  p4-self-natural/      Self-generated text, natural writing condition
    15204.tif             Writer 152 (page 4)
    15304.tif             Writer 153 (page 4)
      .
      .
```

The scans of page 4 (Self-generated text, natural writing condition) have been annotated in ASCII, at the line level. Future work will be directed at lower-level truthing. This can be achieved, possibly, on the basis of a WEB-based tool made available through the openMIND initiative (www.openmind.org).

Chapter 4

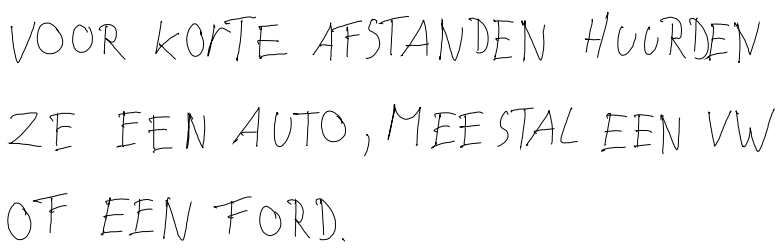
Appendix I. Collection manual for writers (Dutch)

Achtergrond
Welkom bij dit experiment. In opdracht van het Gerechtelijk Laboratorium in Rijswijk doen wij van het NICI (Nijmeegs Instituut voor Cognitie en Informatie) een onderzoek naar de bruikbaarheid van een aantal handschrift herkenningssystemen. Hiervoor hebben we een groot aantal handschriften nodig. En we willen graag ook jouw handschrift hiervoor gebruiken. We garanderen dat de verzamelde handschriften anoniem (onder nummer) bewaard zullen worden. In totaal vragen we je vier teksten te produceren. Hier zul je zo'n 30 tot 45 minuten mee bezig zijn.

Algemene instructies
1) Vul aub op ieder vel in: a) je geboortedatum, b) je huisnummer, c) je geslacht en d) of je links- of rechtshandig bent.
2) Leg onder ieder schrijfvel de blauwe onderlegger.
3) Kom aub niet met de pen buiten de gele markeringen.
4) Schrijf steeds in een normaal tempo en probeer niet overdreven netjes te schrijven.
5) Sla een regel over, daar waar in de tekst lege regels staan aangegeven.
6) Het is niet nodig om alle woorden die in de tekst op één regel staan, ook op één regel te schrijven.

Tekst 1
Allereerst is het de bedoeling dat je de eerste tekst overschrijft, in het handschrift dat je normaal gebruikt.

Tekst 2
Vervolgens is het de bedoeling dat je de tweede tekst in blokletters overschrijft.



Voorbeeld van blokletterschrift

Tekst 3
Hierna willen we je vragen om het derde tekstje zo over te schrijven dat het net lijkt alsof iemand anders het geschreven heeft (een zgn. verdraaid handschrift). Je mag zelf bepalen hoe je dit precies gaat aanpakken (bijv. door veranderen van lettergrootte, schuinheid, lettertype of schriftsoort).

Tekst 4
Als laatste wordt je vriendelijk verzocht om de cartoon op het laatste blaadje in je eigen woorden, in ongeveer zes tot acht regels te beschrijven. Doe dit met volledige zinnen en niet in de vorm van losse steekwoorden. Schrijf hierbij weer in je normale handschrift.

Wij willen je alvast hartelijk bedanken voor je medewerking!!!!

Tekst 1 Schrijf deze tekst over in je normale handschrift

Bob, David en sexy Xantippe sparen postzegels van de landen Egypte, Japan, Algerije, de USA, Holland, Italië, Griekenland en Canada.

!!!!!!!!!!!!!!!!!!!! **LAAT HIER EEN REGEL OPEN** !!!!!!!!!!!!!!!!!!!!

Zij bezochten veilingen en reisden met de KLM. Voor korte afstanden huurden ze een auto, meestal een VW of een Ford.

!!!!!!!!!!!!!!!!!!!! **LAAT HIER EEN REGEL OPEN** !!!!!!!!!!!!!!!!!!!!

De veilingen waren van 7-4-1993 tot 3-5-1993 in New York, Tokyo, Québec, Phoenix, Rome, Parijs, Zürich en Oslo.

!!!!!!!!!!!!!!!!!!!! **LAAT HIER EEN REGEL OPEN** !!!!!!!!!!!!!!!!!!!!

Omdat de veilingen steeds begonnen om 12 uur en je gemiddeld 200 tot 300 kilometer moest rijden, stonden zij steeds om 6.30 uur op en vertrokken om 8 uur uit het hotel.

!!!!!!!!!!!!!!!!!!!! **LAAT HIER EEN REGEL OPEN** !!!!!!!!!!!!!!!!!!!!

Elke dag hadden ze vijfhonderd (f 500,-) gulden nodig. Daarvoor gebruikten ze elke keer een cheque van tweehonderd (f 200,-) en een cheque van driehonderd (f 300,-) gulden. Aan geschenken gaven ze ongeveer honderd gulden (f 100,-) uit.

Tekst2 Schrijf deze tekst over in blokletters

NADAT ZE IN NEW YORK, TOKYO, QUÉBEC, PARIJS, ZÜRICH EN OSLO WAREN GEWEEST, VLOGEN ZE UIT DE USA TERUG MET VLUCHT KL 658 OM 12 UUR.

!!!!!!!!!!!!!!!!!!!! **LAAT HIER EEN REGEL OPEN** !!!!!!!!!!!!!!!!!!!!

ZE KWAMEN AAN IN DUBLIN OM 7 UUR EN IN AMSTERDAM OM 9.40 UUR 'S AVONDS. DE FIAT VAN BOB EN DE VW VAN DAVID STONDEN IN R3 VAN HET PARKEERTERREIN. HIERVOOR MOESTEN ZE HONDERD GULDEN (F 100,-) BETALEN.

Tekst3 Probeer deze tekst zo over te schrijven dat het net lijkt alsof iemand anders het geschreven heeft

Nog dezelfde avond reden ze naar hun vrienden Chris, Emile, Jan, Irene en Henk, nadat ze hun vriendinnen Greta en Maria hadden opgehaald.

!!!!!!!!!!!!!!!!!!!! **LAAT HIER EEN REGEL OPEN** !!!!!!!!!!!!!!!!!!!!

Samen hadden ze vijfhonderd (500) zeldzame postzegels gekocht, Bob driehonderd (300) en David tweehonderd (200).

!!!!!!!!!!!!!!!!!!!! **LAAT HIER EEN REGEL OPEN** !!!!!!!!!!!!!!!!!!!!

De reis was de moeite waard geweest.

Tekst 4 Beschrijf onderstaande cartoon, met een paar zinnen, in je eigen woorden en in je normale handschrift