First UNIPEN Benchmark of On-line Handwriting Recognizers Organized by NIST

CALL FOR DATA (please post)

At the initiative of Technical Committee 11 of the International Association for Pattern Recognition (IAPR), the UNIPEN project was started to stimulate research and development in **on-line handwriting recognition** (e. g. for pen computers and pen communicators). UNIPEN provides a platform of data exchange at the Linguistic Data Consortium (LDC) and is organizing this year a worldwide benchmark under the control of the US National Institute of Standards and Technologies (NIST). **The benchmark is concerned with writer independent recognition of sentences, isolated words and isolated characters of any writing style (handprinted and/or cursive)**. Although UNIPEN will provide, in the future, data for various alphabets, this particular benchmark is limited to **letters and symbols from an English computer keyboard**. The data will be donated by the participants.

Conditions of participation

Participation to the benchmark is open to any individual or institution who provides a sample of handwriting in the UNIPEN format which contains at least 12,000 characters. The data must be of acceptable quality and donated by **October 1, 1994**. The database of donated samples will be available for free to the data donators. Registration material can be obtained by sending email to stan@magi.ncsl.nist.gov or via ftp:

```
ftp ftp.cis.upenn.edu
Name: anonymous
Password: [use your email address]
ftp> cd pub/UNIPEN-pub/documents
ftp> get call-for-data.ps
ftp> quit
```

Organizing committee

Isabelle Guyon, AT&T Bell Laboratories, USA
Lambert Schomaker, Nijmegen Institute for Cognition and Information, The Netherlands
Stan Janet, National Institute of Standards and Technologies, USA
Mark Liberman, Linguistic Data Consortium, University of Pennsylvania, USA
Réjean Plamondon, IAPR, TC11, Ecole Polytechnique de Montréal, Canada

First UNIPEN Benchmark
of On-line Handwriting Recognizers
Organized by NIST

Registration Form

Mail or **fax** (no email) to:

Stan Janet
UNIPEN benchmark
National Institute of Standards and Technology
Building 225, room A-216
Gaithersburg, MD 20899
Phone: +1 (301) 975-2916
Fax: +1 (301) 840-1357
Email: stan@magi.ncsl.nist.gov

Participants must register to be listed on the benchmark mailing list and to get access to the FTP site at LDC.

Participation in the benchmark implies the acceptance of the conditions of data donation and the benchmark protocol. The organizing committee reserves the right of making some changes to the benchmark protocol after receiving the data.

The database which will be constituted both from the samples donated by the participants and the data collected by LDC will be made available for free to the data donators. However, NIST reserves the right of refusing data of too poor quality. Data donated and accepted may or may not be used for the benchmark.

Data donated and accepted give the right to participate in the benchmark but no obligation. Participants can withdraw from the benchmark any time before they release to NIST their recognition results during, the final test. Data which has been donated will not be given back.

Participation in the benchmark implies the donation of a database containing at least one homogeneous set of 12,000 characters (or the equivalent in words or sentences), segmented and labeled at the "text", "word" or "character" level, formatted in the UNIPEN format, and deposited at the FTP site at LDC before October 1st, 1994. Donations of several data sets or of larger data sets are encouraged.

After the benchmark, the training and test data will be distributed by the LDC gratis to its members, and to others on payment of a fee. The benchmark test data will become public domain and will be distributed by anonymous ftp. The results of the benchmark will be published in the open literature.

Name:
Affiliation:
Address:
Phone:
Fax:
Email:
Approximate quantity of data that will be donated (in number of characters):

Date and signature:

First UNIPEN Benchmark
of On-line Handwriting Recognizers
Organized by NIST

Contents

1 Historical Perspective 4

2 Purpose of the Benchmark 5

3 Who Should Participate and Why? 7

4 How to Format the Data? 8

5 Protocol of Data Donation 10

A FTP-site 11

B UNIPEN 1.0 format 12
B.1 Format syntax 12
B.2 Format definition 12

C Data Skeletons 24

D Tips for collecting data 38

E IAPR 45

F LDC 46

G NIST 47

1 Historical Perspective

On-line handwriting recognition addresses the problem of recognizing handwriting from data collected with a sensitive pad which provides discretized pen trajectory information. Contrarily to other pattern recognition fields, such as speech recognition and optical character recognition, no significant progresses have been made, in the past few years, in *on-line handwriting recognition* to make large corpora of training and test data publicly available, and no open competitions have been organized.

The first impulse to UNIPEN was given at the 11th IAPR-IEEE International Conference on Pattern Recognition, in September 1992, by a group of experts, the Technical Committee 11 of the IAPR, Professor Réjean Plamondon chairman. Information on the International Association for Pattern Recognition (IAPR) and the Technical Committee 11 can be found in appendix E. Two IAPR delegates (Isabelle Guyon and Lambert Schomaker) were designated to explore the possibility to create large databases for on-line handwriting recognition research and development.

A small working group was constituted to get the project stated. In May 1993, a nucleus of experts in on-line handwriting recognition (Tetsu Fujisaki (IBM), Ronjon Nag (Lexicus), Sandy Benett (GO/EO), Dick Lyons (Apple), Yves Chauvin (NetID), Dave Reynolds and Dan Flickinger (HP), Isabelle Guyon (AT&T) and Lambert Schomaker (NICI)) layed the foundations of UNIPEN. It was proposed that a common data format would be designed to facilitate data exchange. It was decided that contacts would be made with the Linguistic Data Consortium (LDC) and the National Institute of Standards and Technologies (NIST) to get the data distributed and arbitrate benchmarks.

In summer 1993, the UNIPEN format was designed, incorporating features of the internal formats of several institutions, including IBM, Apple (Tap), Microsoft, Slate (Jot), HP, AT&T, NICI, GO and CIC. The format was then tested independently by the members of working group. A second iteration of test was organized in autumn 1993 to check the changes and additions to the format. In particular, the benchmark protocol was tested. The resulting format was internally used at AT&T and NICI (the home institutions of Isabelle Guyon and Lambert Schomaker) to collect data and benchmark recognizers.

In parallel, at set of tools to parse the format and to browse the data was developed at AT&T and NICI.

In January 1994, the negotiations with LDC and NIST concretized into the organization of the first UNIPEN benchmark. There is now an FTP site at LDC where data and programs can be exchanged and NIST is supervising the benchmark. For information about LDC and NIST, see appendices F and G. For information about the FTP site, see appendix A.

In March 1994, UNIPEN advertized its existence on several electronic mailing lists, resulting in nearly 200 subscriptions to the UNIPEN newsletter.

In June 1994, the instructions for participations to the first UNIPEN benchmark, limited to the Latin alphabet, are released. October 1st, 1994 is the deadline for submitting data. The benchmark will take place in 1995.

The activities of UNIPEN will expand in the future according to the needs and desires of the participants.

2 Purpose of the Benchmark

The three key features of the UNIPEN benchmark are that:

1. the test data will be donated by the participants,
2. the tests of the recognizers will be run by the participants,
3. the statistics on the results will be published at a workshop.

The nature and structure of the benchmark were chosen to serve several goals:

1. Constitute a sizable, quality database which will be available for free to all the data donators and will be later (when the benchmark is over) distributed by LDC.
2. Evaluate the state of the art in on-line handwriting recognition by testing many recognizers in the same conditions on several tasks of various difficulty.
3. Bring together researchers and developers in on-line handwriting recognition from the Academia and the Industry to promote exchanges and collaborations.

The benchmark was also designed to satisfy a number of constraints such as material constraints (rely on voluntary work, use a small budget, get started rapidly):

1. **Data.** We believe that data diversity is an indispensable feature which can be obtained only via the donation of data coming from a large variety of source. Thanks to data donation, we expect to obtain samples of various writing styles, handprint and cursive, isolated characters, isolated words, phrases, sentences, paragraphs, etc. from a large diversity of vocabularies. This will prevent our test set from being biased towards a specific kind of data. If you plan to collect data specifically for the UNIPEN benchmark, please read appendix D.

   Relying on data donations rather than collecting our own data is not justified by material constraints. We estimated that the cost of collecting a "raw" database of 100,000 words is roughly $24000, broken into: 5 data collection stations (PC + disk + pad) at $3000 plus 500 hours of data donators at $15 per hour and 100 hours of data collection supervisor at $15 per hour. This estimate does not include software development, manual segmentation and data cleaning costs. To ensure that a minimum amount of training data will be available to all participants, LDC is financing the collection of a training database of approximately 100,000 words.

   The downside of the data donation paradigm is that there is a risk of constituting an heterogeneous database, each part being too small to be usable by itself. To remedy this problem and ensure some data homogeneity, this first benchmark is limited to the Latin alphabet (plus numbers and some symbols). For further details, see section 4. We anticipate that other data collections and benchmarks along similar lines will be organized in the future for other more complicated tasks and other alphabets.

2. **Test.** The test will be broken into several sections with various levels of difficulty. This will give to everyone the opportunity to achieve excellence on at least part of the test. Note that achieving excellence on your own data sets will not be considered a primary achievement. Average scores will be computed separately on your own data sets and on those offered by others.

Material constraints prevented us from opting for labor intensive solutions such as running recognizers tests "on-site" under the supervision of a benchmark officer, or porting all the recognizers to the same benchmark platform. The last solution would have been anyways impractical to benchmark commercial recognizers which are bound to particular hardware platforms and which source code is not available.

Instead, we decided that the tests will be run by the designers of the recognizers themselves. NIST will provide them with unlabeled data. They will return to NIST the recognition results. NIST will compute the recognition performance. We rely on the integrity of the participants not to attempt labeling the data themselves in order to tune their recognizers by using the test data.

To encourage fair competition, a development test set (with labels) will be available in advance. During the development period, participants will have access by email to the data scoring machine which will return statistics on results formatted in the UNIPEN format. Once the real benchmark test set is released, participants will be under time pressure to produce the results.

The date of the final test has not been fixed yet. Participants will have about one year to work between the release of the training data and development test set and the final test.

3. **Workshop.** The results of the benchmark will be announced by NIST at the workshop organized at the end of the final test. The results will not be anonymous, to avoid technical and legal complications. Participants will have the choice to withdraw any time before the final test is run, in which case their results will not be published.

   The number of participants to the workshop will be limited. Would that become necessary, priority will be given to the donators of large data sets and to those who are willing to publish their results. Otherwise, the workshop will be open to all data donators.

3 Who Should Participate and Why?

You may find many reasons to enter the UNIPEN benchmark, that is, to donate data and eventually to participate in the test, for example:

- You are curious to know what the state-of-the-art is on your own data set.
- You need more training and test data and you would like to use this opportunity to exchange data.
- You are looking for new challenging tasks.
- You think that you will do well in the competition which may secure some funding to pursue your research or will be good advertisement to sell your products.
- You want to use the competition to make contacts with other researchers and identify solutions to your problems.

People frequently ask what kind of data they should give. Is it to the donator's advantage to:

1. give a data set that he is very familiar with, hoping to perform well at least on that set? - Not necessarily: the results on the donator's own data sets will be scored independently.
2. give data that is unreasonably hard? - Not necessarily since everybody will then perform poorly and it will not be possible to evaluate the relative strength of the recognizers.
3. give data that is extremely easy? - Chances are that everybody will do well on easy tasks. However, for practical purposes, it may be important to know what the state-of-the-art is, even on easy tasks (e.g. the recognition of isolated digits).
4. give data that his own recognizer cannot handle? - Possibly: that way he will know how far other groups have already gone on that particular task.
5. give nonsensical data? (e.g. characters that have all the same labels)? - This is not recommended at all.

Compensation for large data donations:

To encourage donations of large quantities of data, LDC will remunerate donations of labeled databases of more than 50,000 characters with a compensation.

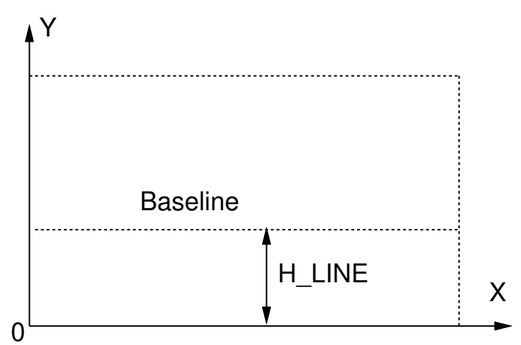
Donators of unsegmented data (i.e. segmentation at the TEXT level only) will receive $2 per 100 characters.

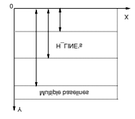
Donators of data which was machine segmented on the basis of bounding boxes or combs (i.e. segmentation at the WORD or CHARACTER level) will receive $5 per 100 characters.

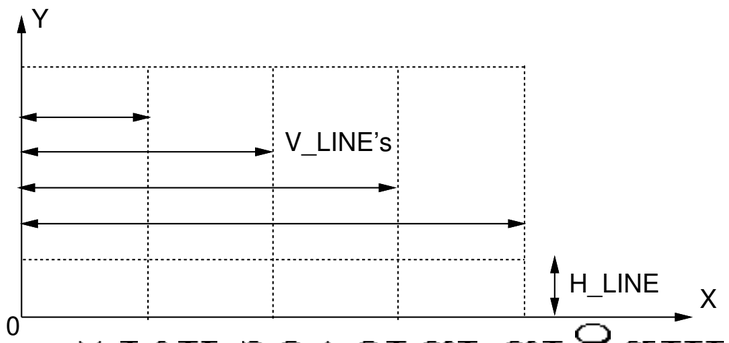
Donators of connected handwriting (run-on or cursive) which was accurately "manually" segmented at the character level will receive $10 per 100 characters.

Donators of data which was completely "manually" reviewed after data collection, sample by sample, by at least two people, to verify the labels, the quality of the samples and the segmentation, will get an additional $5 per 100 characters.

This amount does not intend to cover all the data collection costs, but will help donators to pursue their data collection effort. LDC has allocated $25,000 to be distributed on a first come first serve basis. LDC reserves the right of refusing data of too poor quality.


Y
Baseline
H_LINE
X
0


0
X
Y
H_LAYER
Δλ/λ(Bi) pressure


Y
V_LINE's
H_LINE
X
0

tical guide lines can be specified. In this example, use .- with one

ment s.

.XDIM [N] Width of the bounding bo in piels
using the resolution of the input
deice not that of the display.
MANDATORY in the documentation file.

.YDIM [N] Height of the bounding bo in piels.
MANDATORY in the documentation file.

.HLINE [N] [.] Distance in piels between the bottom
of the bounding bo and horiontal
guidelines such as a baseline.
MANDATORY in the documentation file
if one or seeral baselines are used.

.VLINE [N] [.] Distance in piels between the left
edge of the bounding bo and ertical
guidelines.
MANDATORY in the documentation file
if combs or ertical guidelines are used.

.XPOINTSPERINCH [N] resolution of the data collection
deice 1 inch 2.5 cm. This is the
resolution of the sensor NOT that of
the display.
MANDATORY in the documentation file
or use .XPOINTSPERMM.

.YPOINTSPERINCH [N] y resolution of the data collection
deice.
MANDATORY in the documentation file
or use .YPOINTSPERMM.

.ZPOINTSPERINCH [N] altitude resolution of the data
collection deice.
MANDATORY in the documentation file
or use .ZPOINTSPERMM.

.XPOINTSPERMM [N] resolution of the data collection
deice in SI units.
MANDATORY in the documentation file
or use .XPOINTSPERINCH.

.YPOINTSPERMM [N] y resolution of the data collection
deice.
MANDATORY in the documentation file
or use .YPOINTSPERINCH.

.ZPOINTSPERMM [N] altitude resolution of the data

collection deice.
MANDATORY in the documentation file
or use .ZPOINTSPERINCH.

.POINTSPERGRAM [N] Pressure resolution of the data
collection deice.
MANDATORY in the documentation file
if pressure coordinates are used.

.POINTSPERSECOND [N] Sampling rate
MANDATORY in the documentation file
if time stamps are not proided.

Pen trajectory

Pen trajectory information is encoded with the $.P\_O$ or $.P\_P$ instructions, with repeated sequences of the coordinates declared with $.OO$.

We avoid using the term "stroke" which may refer to a unit of handwriting movement, with respect to a specific generation model. We use instead the term "component" for any continuous trace of the pen recorded by the digitizer. A pen-down component is a trace recorded when the pen is in contact with the surface of the digitizer. A pen-up component is a trace recorded when the pen is at proximity of the digitizer without touching it. Components are not necessarily delimited by pen-lifts and may or may not coincide with strokes.

Thus, the instruction $.P\_O$ does not indicate necessarily that the pen was lifted before. A sure pen-lift is indicated by a $.P\_P$ (an empty component, if no data is recorded when the pen is in the up position).

Since, it is always preferable to provide data which is as raw as possible, if the data collection device does not provide penup/down information, do not try to infer that information, add pressure or time stamp coordinates.

The succession of two $.P\_O$ components not separated by a $.P\_P$ permits indicating a finer segmentation, not based on pen-lifts (for cursive for instance). See also the notation $[A\ [: M]] - [B\ [:]], []$ used in $.SGM$.

The instruction . permits precising the elapsed time between two components when only XY coordinates are recorded and when the sampling rate is regular.

.PENDOWN [N] [.] Pen down component: repeated sequences of
coordinates as defined by .COORD
MANDATORY .

.PENUP [N] [.] Pen up component: same as .PENDOWN but
with the pen not touching the pad
surface.
MANDATORY only if the pen-lift information
is aailable from the pad sensors.

.DT [N] Elapsed time measured when pen
coordinates are elided because the pen

```
                                    was immobile or out of proimity of
                                  the pad sensors.
                                   Very highly recommended.
```

Data annotations

It is important to be able to distinguish between writer dependent and writer independent applications. Thus one must keep track of the identity of the writers across training set and test set. For reasons of confidentiality, it is not always possible to use the full name of the writer as identification. Use the keyword . _ to provide a unique writer identification. We suggest to use first name and last name initials and a four digit extension phone number (e.g. G_7523). The person who is working at the data source and identified by $.AA\_SO$ is responsible for the book keeping of the writer identifications. Note that an ID consisting of a machine name and a date is not adequate since the same writer may come back to donate data at a later point in time. It is very highly recommended to provide all possible informations about the writer, even if they are not mandatory.

```
.DATE [N] [N] [N]              Date stamp: month day year.

.STYLE [R]                   PRINTED CURSIVE or MIXED.
                             MANDATORY .

.WRITERID [S]                  Unique writer identification.
                               MANDATORY .

.COUNTRY [S]                 Country of origin.

.HAND [R]                    Writer hand: L for left R for right.

.AGE [N]                     Writer age in years.

.SEX [R]                     Writer se: M for male F for female.

.SKILL [R]                   Skill of writer familiarity with
                             input deice: BAD OK or GOOD.

.WRITERINFO [F]              Misc. information about writer.
```

Data sets and implicit component numbering

The database is divided into one or several data sets. Within a set, components starting either with $.P\_O$ or $.P\_P$ are implicitly numbered, starting from zero. Empty components are NOT counted.

For this benchmark, **sets will coincide with data files**, each one containing data from one writer. The instruction $.SA\_S$ is not needed. There will be one single component numbering for each file, starting from zero.

```
.STARTSET [S]          Start a new set the argument is
                      the set name.
                      The component counter is reset to ero.
                     In the absence of .STARTSET the
                    component counter is automatically reset
                    to ero at the beginning of each file
                   and the set name is the file name.

.STARTBOX [.]        Instruction to the data data isualisation
                    program browser to clear the data bounding bo.
```

Segmentation and labeling

Component numbers are used by the $.SGM$ instruction to delineate sentences, words, characters, etc. with a $[] - [], []$ expression. To indicate a finer segmentation containing broken components, one can use the more complex notation $[A\ [: M]] - [B\ [:]], []$ (see reserved string glossary).

In addition to the hierarchical segmentation declared in $.AY$, undeclared segmentation information may be introduced to encode various kind of break points such as chapters, pages, blocks or lines (e.g. $.SGM\ \ PAG$ 203 – 621).

```
.SEGMENT [S] [R] [R] [L]       Type of segment its delineation
                              quality and label.
                               .SEGMENT is  MANDATORY  for the
                             benchmark it is needed to proide the
                            transcripts or labels.
                             -> First argument: type of segment
                           such as the ones declared in
                           .HIERARCHY e. g. TEXT WORD
                        CHARACTER etc..
                       -> Second argument: segment
                      delineation by a A[:M]]-[B[:N]][C]
                     epression see resered string
                   glossary. Components are numbered
                  starting from ero in order of
                 apparition in the current set
                 which coincides for the benchmark
                with the entire data file of one
                writer. Empty components are
               NOT counted.
                -> Third argument: quality leel
            BAD for illegible OK for regular
          GOOD for superior  for unknown.
        -> Fourth argument: transcript or label
      Sentence word character etc.
```

Reserved string glossary

`[N]-[N][N]` Compact representation of a list of numbers used by .LEXICONSET .SEGMENT .RECTIME .RECLABELS .RECSCORES. The list: 2 3 4 5 15 9 50 51 52 53 54 55 would be represented as: 2-515950-55. Commas allow non contiguous numbers useful for segmentation of delayed strokes i dots and t bars. NO SPACES are allowed in the notation.

`[A[:M]]-[B[:N]][C]` More fleible representation to delineate segments of data by breaking components. AB and C are component numbers M and N are point numbers in the component. Both components and points are 0-base. L defaults to ero and M to the last point in the component. Thus the [N]-[N][N] epressions are special cases of a [A[:M]]-[B[:N]][C] epression. The eample 1:40-356:0-6:12 delineates component 1 from point 40 to the end all of component 2 3 and 5 and component 6 from the beginning to point 12. NO SPACES are allowed in the notation.

`X` X position of the pen on the pad surface in units of X gien by .XPOINTSPERINCH

`Y` Y position of the pen on the pad surface in units of Y gien by .YPOINTSPERINCH

`T` Time in MILLISECONDS.

`P` Pressure in units of P gien by .UNITSPERGRAM. Use preferably a linearied pressure in 1000 units per gram of force and calibrate the ero as the threshold of pen reaching the pad surface. Negatie pressures

	can account for remaining non-linearities and hysteresis.
Z	Altitude aboe the pad surface in units of Z gien by .ZPOINTSPERINCH.
BUTTON	Barrel button states: 0 1 ...
RHO	Rotational angle of the stylus measured in degrees from some nominal orientation of the stylus e. g. barrel button on top. The angle increases with clockwise rotation as seen from the rear end of the stylus.
THETA	XY angle of the stylus measured in degrees increasing from the X ais in the counter-clockwise direction.
PHI	Z angle of the stylus increasing from the pad surface in the positie Z direction.
L	Left handed.
R	Right handed.
M	Male.
F	Female.
BAD	Unskilled writer illegible writing.
OK	Aerage quality writing unambiguously legible.
GOOD	Superior quality writing most recogniers should get it.
	Unknown.
PRINTED	Printed handwriting style.
CURSIVE	Cursie handwriting style.
MIXED	Mied printed and cursie handwriting style.

C Data Skeletons

The following data skeletons designed for the UNIPEN benchmark can be retrieved at the FTP site of LDC from directory:

pub/UNIPEN/databases/examples/org/

In the example directory tree, substitute to "org" your three letter organization identification received upon registration. Should you donate several databases, create sub-directories orgX, where X is a number between 0 and 9. Each database should be homogeneous and contain at least 12,000 characters, or the equivalent in words or sentences. Within a sub-directory orgX, each file must contain data from only one writer.

```
                                    org/

                    ---------------------------------------------

              README   documents/ include/ org0/   org1/    org2/

          documents: A directory containing releant documents including engineering
                   drawings of the data collection forms papers or
                  technical reports published about the database or performance
                  results using the database. All documents should be in
                  PostScripttm format.

                include:   A directory which should contain all include files.

                org0:      Eample of isolated character data set.
               org1:      Eample of isolated word data set.
              org2:      Eample of sentence data set.


              File: org/include/org0.doc

            .VERSION 1.0
            .COMMENT
      SKELETON OF DOCUMENTATION FILE FOR ISOLATED CHARACTER DATABASE

ull name of your organiation.

his database where org is the 3 letter
on which was assigned to you upon
e name of this file which is used

se do not change it and conform to
```

```
                                    .DATASOURCE    organiationname
                                  .DATAID        org0

                                 .COMMENT Use the .ALPHABET specified the same for all the databases of the
                                        benchmark. It should be a superset of the character labels that
                                       you are using. Optionally use .ALPHABETFREQ to specify character
                                      frequencies by replacing the 1s by appropriate alues.

                              .ALPHABET       "0" "1" "2" "3" "4" "5" "6" "7" "8" "9"
                                              "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" "" "w" "" "y" ""
                                             "" "" "" "" "" "" "" "" "" "" "" "-" ""
                                      "" "" "" "/" "" "" "" "[" "]" """ ":" "" ">"
                                 "" "." "" "" "" "" " "

.ALPHABETFREQ  1 1 1 1 1 1 1 1 1 1
               1 1 1 1 1 1 1 1 1 1 1 1 1
               1 1 1 1 1 1 1 1 1 1 1 1 1
               1 1 1 1 1 1 1 1 1 1 1 1 1
               1 1 1 1 1 1 1 1 1 1 1 1 1
               1 1 1 1 1 1 1 1 1 1 1 1 1
               1 1 1 1 1 1 1

                              .COMMENT Use the .HIERARCHY CHARACTER for isolated characters as specified.
                                       This means that only the .SEGMENT type CHARACTER is allowed
                                       in data files which include this file. Character labels may be any
                                       label from .ALPHABET.

                              .HIERARCHY      CHARACTER

                              .COMMENT
                           DOCUMENTATION


    Please fill up the blanks and add any other releant information.

  .DATACONTACT
          Name:
          Affiliation:
          Address:
          Phone:
          Fa:
         Email:

.SETUP   Site:
```

```
                    Time:
                    Writer population:
                    Writer motiation: paid/olunteer/forced
                  Writer physical position:
                  Instructions gien to the writer:
                 Duration of one session:
                 Recognier feedback: yes/no
                Form layout:

      .DATAINFO
              Set type: training or test

      Alphabet: English keyboard alphanumerics and symbols
                from the ASCII set
      Leicon:
     Label distribution:
     Quantity:
     Quality: cleaning/no cleaning
     Number of Writers:
   Writing Style: printed/cursie/mied
 Segmentation:

.PAD     Machine name:
         Brand:
         Type:
         Serial Nr.:
         Sensor:
         Pen:
         Drier:
        Sampling mode:
        Sampling rate:
        Resolution:
        Accuracy:
        Width:
        Height:
        Display:
        Inking:


.COMMENT
DATA LAYOUT


      Please replace numbers by adequate alues. .XDIM and .YDIM define
    the bounding bo. .HLINE is the baseline which may be optionally
  specified. All characters must fit in the same bounding bo. The
 UNIPEN coordinate system must be strictly obsered 0 at the lower
left corner X ais pointing to the right Y ais pointing to the
```

```
                                                                                                    top.

                                                                                                .XDIM    75
                                                                                               .YDIM   160
                                                                                              .HLINE   75

                                                                                             .COMMENT
                                                                                      UNIT SYSTEM


                                                                  Please replace 000 by adequate alues. .XPOINTSPERINCH and
                                                                .YPOINTSPERINCH may be replaced by .XPOINTSPERMM and
                                                              .YPOINTSPERMM. .COORD must hae at least X and Y but may
                                                                include too P pressure and/or T time the altitude aboe
                                                                the pad Z or een angle information RHO THETA PHI
                                                                if aailable see UNIPEN documentation. If you use pressure
                                                                information conform to the UNIPEN documentation and
                                                                use .POINTSPERGRAM to precise the resolution.

                                                .XPOINTSPERINCH 000
                                              .YPOINTSPERINCH 000
                                            .POINTSPERSECOND 000
                                          .COORD             X Y


                                          File: org/include/org1.doc
```

This file is similar to org/include/org0.doc, except that org0 is replaced by org1 and . AY AA by:

```
.HIERARCHY      WORD

.COMMENT Use .HIERARCHY WORD for isolated words that are not
         segmented into characters and .HIERARCHY WORD CHARACTER
         if segments of both .SEGMENT types WORD and CHARACTER appear
         in the data file in which this file is included.
         Words may contain any characters from .ALPHABET ecept
        white space.

File: org/include/org2.doc
```

This file is also similar to org/include/org0.doc ecept that org0 is replaced by org2 and .HIERARCHY CHARACTER by:

```
.HIERARCHY      TEXT WORD

.COMMENT Use .HIERARCHY TEXT for unsegmented tet.
       Use .HIERARCHY TEXT WORD if both .SEGMENT types TEXT and
```

```
WORD appear in the data file in which this file is included.
Use .HIERARCHY TEXT WORD CHARACTER if all three types are
used. The tet labels may contain any characters from .ALPHABET.
If there are restrictions on the character set please
specify them with .ALPHABETSET in the data file directly.
```

Notice also that multiple baselines are defined in that file with:

```
.HLINE  450 1900 2300
```

```
File: org/include/phrase.le

.VERSION 1.0
.COMMENT
SKELETON OF UNSEGMENTED LEXICON

     We leae the option to proide an unsegmeted leicon
  e.g. a list of sentences or phrases from which the data
 was drawn or an entire tet. It is also possible to proide
a segmented leicon a list of words for a tet database.
                  organiation-name is the full name of your organiation.

                phrase is here the name of that leicon also the name of the file
              used to identify the leicon in .INCLUDE.

      .LEXICONSOURCE organiation-name
    .LEXICONID     phrase

   .COMMENT Please fill up the blanks and add any other releant information

  .LEXICONCONTACT
          Name:
          Affiliation:
          Address:
          Phone:
          Fa:
         Email:

.LEXICONINFO
        Source:
        Purpose:
        Language:
        Label distribution:

.COMMENT Please proide all labels between "". Do not forget that
       the label quote  " and the label backslash  .
```

```
                                                    .LEXICON
                                                       "Thomas N. Zinfindale 6877 Greensburg"
                                                      "Drie Phoeni MN 34132. Hanna Q."
                                                  "Allen 9887 County Drie Phoeni ME"
                                               "39153. Bruce C. Prescott 8138 South"
                                             "Court Columbus UT 56029. Doris O."
                                          "Daies 8152 Bluedale Drie"
                                      "Knoille SC 12277. Sally Y."
                                   "Hinricks 6837 Lake Drie Boise RI"
                               "81516. Hank I. Samuals 2685 Lake"
                             "Aenue Madison WA 03915. Frank O."
                         "Allen 6216 Bluedale Street Topeka"
                    "South Terrace Boise KY 29445."


                      File: org/include/word.le


                  .VERSION 1.0
                  .COMMENT
SKELETON OF WORD LEXICON

                    organiation-name is the full name of your organiation.

                   word is here the name of that leicon also the name of the file
                 used to identify the leicon in .INCLUDE.

      .LEXICONSOURCE organiation-name
    .LEXICONID     word

   .COMMENT Please fill up the blanks and add any other releant information

  .LEXICONCONTACT
          Name:
          Affiliation:
          Address:
          Phone:
          Fa:
         Email:

.LEXICONINFO
        Source:
        Purpose:
        Language:
        Label distribution:
```

```
.COMMENT Please proide all labels between "". Do not forget that
        the label quote  " and the label backslash  .
  Optionally use .LEXICONFREQ to proide word frequency
 information as computed for instance from a corpus of tet.

.LEXICON
"the"
"of"
"to"
"and"
"a"
"in"
"said"
"that"
"for"
"is"
"was"
"on"
"with"
"by"
"as"
"he"
 "it"

 .LEXICONFREQ
11000000
5824125
4958348
4683121
4167577
4053021
1924978
1889885
1796491
1557596
1496299
1362125
1137819
1060304
977732
965713
920947
```

```
                                                                    .VERSION 1.0
                                                                    .COMMENT
                                    SKELETON OF DATA FILE FOR ISOLATED CHARACTER DATABASE


      Include the documentation file mandatory.
    Included files reside in the include/ directory and
    shall come with no path. No file should be included in an
    included file.

    arbi is the writer identifier also the name of this file.
  There should be one file for each writer.

E org0.doc
ID arbi

Replace eros or question marks by appropriate alues
e UNIPEN documentation. All these entries are optional

0 00



 .WRITERINFO with any other information that has not
itemied.



age:
ges:
th:
ed:
idence:
rience:


ples of characters borrowed from samples of the IBM
 X and Y coordinates were used: no time coordinate
gular sampling rate is assumed no .DT no pressure
pen up strokes no angle information. Consult
```

the UNIPEN documentation to encode more comple eamples.

```
   .SEGMENT CHARACTER 0 OK "P"
   .PENDOWN
  23 114
  22 116
  22 118
  etc.
  23 97
  27 93
  32 90
  .PENUP

 .SEGMENT CHARACTER 1 OK "e"
 .PENDOWN
19 88
18 89
17 89
etc.
57 78
59 77
59 76
   .PENUP

   .SEGMENT CHARACTER 2-3 OK "j"
   .PENDOWN
  34 100
  33 100
  33 100
  etc.
  9 53
  10 53
  13 54
  .PENUP
 .PENDOWN
27 116
27 116
27 116
etc.
32 108
34 106
34 105
.PENUP
```

```
                                                                                .VERSION 1.0
                                                                                .COMMENT
                                                                SKELETON OF DATA FILE FOR WORD DATABASE


                    Include the documentation file mandatory and
                   the leicon optional.
                Included files reside in the include/ directory and
                shall come with no path. No file should be included in an
                included file.

                tobito is the writer identifier also the name of this file.
               There should be one file for each writer.

               Replace .STYLE PRINTED by the appropriate information
               PRINTED CURSIVE or MIXED.

        .INCLUDE   org1.doc
        .INCLUDE   word.le
                .WRITERID tobito
               .STYLE    PRINTED

               .COMMENT Replace eros or question marks by appropriate alues
                      see UNIPEN documentation. All these entries are optional

            .DATE    00 00 00
            .HAND
           .SKILL
          .COUNTRY
         .AGE     00
         .SEX

        .COMMENT Complete .WRITERINFO with any other information that has not
                yet been itemied.

      .WRITERINFO
              First name:
              Last name:
              Natie language:
             Other languages:
             Place of Birth:
             Where educated:
             Place of Residence:
             Computer eperience:
            Occupation:
```

Vision:

```
.COMMENT These are eamples of characters borrowed from samples of the HP
        database. Only X and Y coordinates were used: no time coordinate
        therefore a regular sampling rate is assumed no pressure
     coordinate and no pen up strokes no angle information. Consult
   the UNIPEN documentation to encode more comple eamples.

    .SEGMENT WORD 0-11  "Tattersall"
   .PENDOWN
  123 122
  123 122
  123 122
  etc.
  117 4
  117 4
  120 5
  .PENUP
 .DT 100
 .PENDOWN
297 151
   296 151
   etc.
   31 79
   17 75
   10 72
   .PENUP
  .DT 100
  .PENDOWN
 240 53
 233 56
 224 56
 etc.
 240 11
 243 5
 247 0
  .PENUP
.DT 100
.PENDOWN
307 109
308 100
311 81
etc.

etc.
```

```
                                                                                .VERSION 1.0
                                                                                .COMMENT
                                                        SKELETON OF DATA FILE FOR TEXT DATABASE


                  Include the documentation file mandatory and
                 the leicon optional.
              Included files reside in the include/ directory and
              shall come with no path. No file should be included in an
              included file.

              camaru is the writer identifier also the name of this file.
             There should be one file for each writer.

             Replace .STYLE MIXED by the appropriate information
             PRINTED CURSIVE or MIXED.

  .INCLUDE   org2.doc
  .INCLUDE   phrase.le
 .WRITERID camaru
 .STYLE     MIXED

 .COMMENT Replace eros or question marks by appropriate alues
 see UNIPEN documentation. All these entries are optional

 .DATE    00 00 00
 .HAND
 .SKILL
 .COUNTRY
 .AGE     00
 .SEX

 .COMMENT Complete .WRITERINFO with any other information that has not
 yet been itemied.

.WRITERINFO
 First name:
 Last name:
 Natie language:
 Other languages:
 Place of Birth:
 Where educated:
 Place of Residence:
 Computer eperience:
 Occupation:
```

Vision:

```
.COMMENT These are eamples of characters borrowed from samples of the ATT
        database. Only X and Y coordinates were used: no time coordinate
        therefore a regular sampling rate is assumed no pressure
     coordinate and no pen up strokes no angle information. Consult
    the UNIPEN documentation to encode more comple eamples.

    In this eample .STARTBOX is used but this is optional.
  Also .SEGMENT PAGE proides additional optional segmentation
 information. The only segments that will be taken into account
 for the benchmark are TEXT WORD or CHARACTER or a selection
of those as precised in .HIERARCHY.

    .STARTBOX
   .SEGMENT PAGE          0-58
   .SEGMENT TEXT          0-29  "that nothing more happened "
  .SEGMENT WORD          0-6  "that"
 .PENDOWN
707 2417
707 2424
706 2429
    etc.
    590 2319
    590 2319
    590 2319
     .PENUP
    .DT 151
    .PENDOWN
   588 2377
   586 2377
   586 2377
   etc.

   etc.

    .PENUP
   .SEGMENT WORD           7-15  "nothing"
  .PENDOWN
1778 2366
1778 2366
1776 2367
etc.

etc.

.PENUP
.SEGMENT WORD          16-19  "more"
```

```
        .PENDOWN
        3048 2355
        3046 2357
        3045 2358
        etc.

        etc.

        .PENUP
       .SEGMENT WORD           20-28  "happened"
      .PENDOWN
     417 2019
     417 2019
     417 2021
     etc.

     etc.

     .PENUP
    .SEGMENT TEXT          30-58  "that nothing more happened "
  .PENDOWN
347 525
     347 532
     347 537
     etc.

     etc.

     .PENUP

    .STARTBOX
   .SEGMENT PAGE           59-102
   .SEGMENT TEXT           59-80  "she decided on going"
  .SEGMENT WORD          59-61  "she"
 .PENDOWN
325 2328
325 2329
325 2330
etc.

etc.
```

D Tips for collecting data

If you are planning to collect data for the UNIPEN benchmark, we invite you to read these recommendations which might save you a lot of trouble and ensure a better quality of the data:

Golden rules

- Record all informations about recording conditions, hardware, software and writer background at the time of the data collection. Annotations made *a oteriori* have higher chances of being wrong or incomplete.
- Collect data rich in information: with the highest possible spacial resolution and sampling rate, with as little preprocessing or filtering as possible, with time stamps, pressure, "a proximity" strokes if available.
- eep the instructions to data donators simple. Before starting the data collection, test your setup on "naive" subjects that have not been part of the design team.
- Before collecting large amounts of data, test your entire system by collecting a small sample. Run it through the UNIPEN parser and visualize some examples with a UNIPEN data visualization program (browser). Such tools are available to you at the FTP site of LDC.

Hardware

Use the best possible hardware you can afford. For instance, use pads which display the "electronic ink" rather than opaque tablets.

The Wacom PL 100- (integrated digitizer and LCD display) is among the best pads available:

- display:GA 640x480 16 grey-level backlit STN LCD
- digitizer: .05 mm/point maximum resolution, 192x144 mm active area
- maximum sample rate: 205 points per second
- communications: unit comes with a PC-AT bus i/f card that handles video and serial i/o to the tablet.
- approximate price: $1500.00
- delivery time: in stock generally
- order from: WACOM CORP ancouver, Washington, USA (1-800-922-6635)

We also recommend the TS-5 ideoTablet:

- display: GA 640x480 64 grey-level backlit STN LCD
- digitizer: .025 mm/point maximum resolution, 196x148 mm active area
- maximum sample rate: 135 points per second
- price: $1495.00
- delivery time: 2-3 weeks
- order from: URTA Phoenix, Arizona, USA (1-800-445-8782)

Avoid Calcomp 2300, 2500 opaque tablets (random noise and high tilt sensitivity) and cheap ($\leq$ $500) tablets intended for drawing rather than for handwriting input.

In general, beware of large screen format capacitive/resistive "membrane" digitizer devices. Users tend to rest parts of the hand/fingers on the writing surface at times, which can lead to spurious information in the tablet data stream. Electro-magnetic digitizers (ie., Wacom/urta) are a better choice.

Some digitizer units use pens that require batteries to power them. As the batteries discharge, pen performance can become erratic without warning... Moral: keep an eye on the batteries used by the pen device.

Driver

There is a choice of commercial drivers which come with the notepad computer or the operating system. A widely distributed software is Microsoft Windows for Pen Computing 1.0. We recommend to:

- use the WFP SD function GetPenHwData for data closest to the tablet: maximum sample rate, no unwanted preprocessing.
- avoid using coordinates generated by WM_MOUSE_MOE events. It is better to collect the original "raw" data and it is always possible to revert *a oteriori* to "mouse-type" spatial sampling.

You may also want to write your own driver to read points coming from the pad. In this case, make sure that you are not dropping points or loosing track of the time stamps. It is indeed very well possible to produce drivers which are better than the ones provided commercially! Note that design considerations for some tablet drivers originate from the crosshair puck used in CAD, where the requirements are different from handwriting (speed is less important in CAD).

General driver requirements:

- Create an interrupt-based driver, do not use polling!
- There are three types of basic functions:

  INIT_SETTINGS, (send characters to the tablet to select dataformat, sampling mode, sampling rate, resolution etc.)

  START_SAMPLING, (send command to start the sampling)

  STOP_SAMPLING. (send command to stop the sampling)

  Allow for the checking and controlling of at least the sampling rate parameter.

  Make sure the storing of coordinates has a far higher priority than the graphical inking, but redesign your driver if there is a noticeable inking delay. Some writers get more upset than others about inking delay. The inking should be (almost) immediate. Try a faster computer (486 instead of 386, Quadra instead of MAC II) if it does not work.

- Use the highest possible sampling rate (it is always possible to perform resampling at a lower rate in software), and not less than 100 points per second.
- Reduce as much as possible or eliminate completely smoothing and low-pass filters: the details of the dynamics of the pen may be irremediably lost. It has been observed with a commercial driver that, for fast writers, a capital B and a capital D are undistinguishable. Also, prevent the use of dehooking or other preprocessing heuristics. This functionality is part of a recognition system, not of the driver.
- Document the features of your driver in *.PA*. In particular, indicate the host machine, the operating system, the estimated current load, the sampling mode (interrupt driven or polling), the number and size of the buffers on the signal path from the tablet to the disk.

Annotations

The data should be annotated with informations about recording conditions, hardware, software, and writer background (see UNIPEN documentation and examples for details). Informations should be recorded at the time of the data collection. Annotations a posteriori have a higher chance of being wrong or incomplete. Design a computer survey form which will be filled at each data collection session. Record the date, the machine and pad serial numbers (useful for tracking hardware failures).

Examples of good/bad annotations (thanks to all who provided examples):

```
-----------------------------------------------------------------------------
.PAD
         Machine name:      WACOM PL100-V
         Brand:             Wacom
         Type:              PL-100V
         Serial Nr.:        2JPJB0044

         Sensor:            Electromagnetic wireless pen
        Pen:               Untethered Pen tip switch and side button.
                          Tip switch used for PenUp/PenDown determination.
       Drier:            Custom-made routine calling GetPenHwData under
                       MS Windows for Pen Computing V1.0 PENWIN.DLL
 Sampling mode:     Equidistant in time during PenUp  PenDown
                                    continuous sampling
                        Sample rate:       100            [H]
                       Resolution:        0.02           [mm/unit]
                       Accuracy:          0.1            [mm]
                       Width:             192            [mm]
                       Height:            143            [mm]

                       Display:           Wacom PL-100V backlit LCD screen VGA 640480
                                     for the writer and a 15" CRT monitor for the
                                     superisor. The writer only looked at the PL-100V
                 Inking:            Graphically emulated black on white 1 piel thick.

      ---------------------------------------------------------------------------

        .PAD sub fields and their alues as in receied files.

    Machine name:   Wacom SD-510C
    Machine name:   Scriptel Penwriter

    Machine name:   Wacom HD-648A
    Machine name:   IBM ThinkPad 700T
    Machine name:   Calcomp 2500
    Machine name:   NCR 3125 pen computer
    Machine name:   Philips Adanced Interactie Display PAID prototype "notebook"
Machine name:   CIC-MacHandwriter CalComp digitier c/to a Macintosh IICI
```

Note: a distinction between brand/type/serial number is necessary see aboe.

Display: Apple monitor
Matri 640 480 points
black ink on white background
Display: backlit liquid crystal display LCD
Matri 640 480 points
VGA standard black on white
Display: translucent type blue mode STN with backlight
Matri 640 480 points
Display: LCD with backlight
Matri 640480 points VGA compatible
Display: VAXstation 3100 15" monochrome screen for inspection
Display: lcd display

VGA resolution 640480
black ink on white background
Display: 640480 VGA 16 gray leels
Display: Macintoshs screen.
Resolution: 72 piels per inch.
Ink: black on white 1 piel thick.

Note: physical sie resolution and inking information are needed.
Thats why there is an Inking: subfield proposed see aboe.

Pen: Untethered pen with tip switch
Pen: Tethered pen with side button
Pen: Untethered pen with side button
Pen: Standard non-stroke with tip and barrel switches
Pen: Ballpoint refill pen with home-made aial pen force transducer

Note: also tell what determines your PenUp/PenDown
tip switch or side button etc

Sensor: Electromagnetic gie and receie action
Sensor: Philips confidential
Sensor: Electromagnetic resonance sensor
Sensor: Resistie
Sensor: Capacitie

Note: also tell whether the pen is wireless.

```
                    ---------------------------------------------------------------------------

                    Drier:         Home made
                   this is too ague
                Drier:         MS-mouse-like
               better but not too informatie
           Drier:         UCL-made MAC drier and interface.
                         Time-equidistant sampling between
                         .PENDOWN and .PENUP.
       still better

      Drier: subfield is ery important and says een more about the data than the
   Machine name: Brand: or Type: subfields.
  ---------------------------------------------------------------------------

  Data donators
```

Try to balance the writer population left/right handed and female/male and collect samples from many different writers. Since many factors that are not itemized by special UNIPEN keywords may inuence the handwriting quality, keep a detailed description of the recording conditions in $.S\ P$ and $.\ \_FO$, including:

- the writer's familiarity with the data collection device
- whether the writer was paid or volunteered
- the duration of the data collection session
- the position of the writer (standing, sitting at a desk)
- vision impairments
- the writer's familiarity with the language

In most data collections, you will want the "built-in" recognizer to be "turned off". Writers do adapt quite fast to the recognizer and keeping the recognizer active will bias the data. If this is your intention, please clearly indicate it in the $.S\ P$.

The instructions to data donators should be short and simple: it should be "obvious" what to do. Before starting to use your data collection software, test it on naive potential donators who have not been part of the design team. Avoid imposing counter intuitive constraints on the writer which are likely to introduce errors in the data (e.g. avoid "Do not try to correct any mistakes", "Do not wrap text from one line to the next"). The software should be robust with respect to donators' mistakes by either recovering from them or warning and asking to rewrite (e.g. if the writer goes out of the bounding box). Let data donators get familiar with the device before starting to record.

REMINDER: You are responsible for the book-keeping of writer identifications. Each writer should have a unique $.\ \_$ so that it is possible to track whether data from the same writer is both in the training set and in the test set. An ID consisting of a machine name and a date is not adequate since the same writer may come back to donate data at a later point in time.

Data layout to facilitate segmentation

Every effort should be made to minimize or eliminate labeling errors *at the orce* in order to avoid the painful data cleaning process. The challenge is to make sure that the right ink is attributed to the right segment.

If your pad is equipped with a display, make good use of it:

- eep the page simple with as few buttons, text, boxes as possible. Just make a help button which opens an explanation screen and one clear button for each writing area so that the writer can selectively erase and rewrite bad entries.
- Give feedback to the writer. For instance, give warnings when the pen goes out of the bounding box (beep + ashing light). Then erase whatever was in that box to force the writer to rewrite by respecting the bounding box.

General recommendations:

- Separate clearly the writing boxes by wide dark margins.
- Preferably always use a baseline in each writing field.
- Make sure that there is enough space in each writing box to write the entire truth label string.
- Make sure that the geometry of the guidelines or combs is comfortable for the writer.
- Each data item corresponding to one truth label string should have its own writing box. The label should be associated without ambiguity to that box (e.g. in a corner of the box itself).
- Ambiguous truth labels should be explained (e.g. ' = quote and , = comma)
- Collect the ink from the entire screen. When the page is totally finished, proceed to segment it to attribute strokes to their legal writing area. This should take care of the problem of delayed corrections.
- eep a record in UNIPEN format of the layout geometry, including the position of baselines, boxes and combs.
- eep an engineering drawing of the layout geometry with dimensions expressed in the same units as the UNIPEN records.

Truth labels

Report what your intention was in choosing the truth labels in $.AA\_FO$. Here are general guidelines which may be violated depending on the target application:

- Select truth labels that cover the entire character set and try to balance character frequencies.
- Select a variety of sources which present different difficulties (e. g. literature texts, advertisements, scientific abstracts, computer programs, computer manual pages, daily news, addresses).

- Do not always repeat the same set of words/sentences for each writer but rather draw fresh words/sentences to get more coverage.
- If you collect isolated characters, do not present them always in the same order.

Data cleaning

If, in spite of all your efforts, your database still contains errors, proceed to clean it with CAUTION:

- It is harmful to loose the original data labeling. It is better to store several sets of .SEGMENT entries for the same data file. To do so, if "zarbi.dat" is your data file, create several files "zarbi.sg0", "zarbi.sg1"', etc. containing the same header information as the data file, but no data, just the segmentation information.
- Always keep a file with the original labels in a "zarbi.sg0" file.
- Do not eliminate bad patterns, mark them with BAD in the quality field of $.SGM$.

- If you use several judges to clean the data, keep separate records of their decisions by creating separate files "zarbi.sg$n$" files.

- Document how you cleaned the data in $.AA\_FO$. In particular, say if you used your own recognizer to help you clean the data by pointing out suspicious patterns.

APR

IAPR is an international association of non-profit scientific or professional organizations (being national, multinational or international in scope) concerned with pattern recognition, computer vision and image processing in a broad sense.

The major event in the IAPR calendar is the biennial International Conference which attracts scientists, researchers and practitioners in the field from all over the world; here they are able to listen, to learn, to educate and to exchange ideas with their colleagues.

IAPR has established a number of Technical Committees (working groups with only a single level technical specialities and to further the objectives of the Association. Each committee has a distinct purpose of scope and is under the responsibility of a chairman appointed by the IAPR President. These committees organize regular or special meetings and workshops as well as collaborative activities with other organizations.

The goal of Technical Committee 11, Professor Rejean Plamondon, Chairman, is to promote the exchange of information on research projects dealing with text processing using both static and dynamic methods for printed or handwritten document understanding. The major areas of research concerned by this committee are: on-line and off-line character recognition, document processing and understanding, cursive script recognition, automatic signature verification, modeling of handwriting generation and perception.

Professor Rejéan Plamondon
Ecole Polytechnique de Montreal
C.P.