# DATA COLLECTION FOR HANDWRITING CORPUS CREATION IN INDIC SCRIPTS

*Mudit Agrawal, Ajay S Bhaskarabhatla and Sriganesh Madhvanath*

Hewlett-Packard Laboratories
Bangalore, India
{mudit.a@hp.com, ajay.bhaskarabhatla@edi.gatech.edu, srig@hp.com}

## ABSTRACT

*Linguistic resources such as annotated corpora are critical for the development of language technologies such as speech and handwriting recognition. This paper describes efforts at HP Labs, Bangalore, to create datasets for the design and development of Online Handwriting Recognition (HWR) algorithms for Indic scripts. "Online" in the context of HWR refers to the fact that handwriting is captured as a stream of points using an appropriate pen position sensor (often called a digitizer), rather than as a bitmap (image). In this paper, we focus on some of the issues to be addressed in handwriting data collection - the design of data to be collected, the recruitment of writers, data collection methodology and process, and relevant tools. We discuss these issues in the context of our own efforts to create handwriting corpora for the Tamil script.*

## 1. INTRODUCTION

Despite being used by more than a billion people all around the world, most of the Indic languages and scripts have seen relatively little targeted research in human language technologies such as speech recognition, text to speech synthesis etc. The average Indian citizen is not computer-savvy and often knows only his or her native language and script. Handwriting continues to play a very important role in all spheres of life - from day to day transactions in businesses to personal communications. In such a scenario, technology for online handwriting recognition (HWR) finds a lot of potential applications that extend the reach of IT to the common man.

There have been a few efforts towards creating HWR engines for Indic scripts such as Tamil, Telugu, Devanagari and Kannada. However, the lack of significant linguistic resources in standard data formats had proved an obstacle to more widespread research into HWR technology for Indic scripts. In this paper we describe our own efforts in compiling handwriting datasets for Online HWR that can support our own research as well as benefit the research community. Here "online" refers to the fact that handwriting is captured as digital ink - a stream of (x,y) points, using an appropriate pen position sensor (often called a digitizer), rather than as a bitmap (image). It should be mentioned that such datasets would also benefit research in handwritten document analysis, writer identification, script identification, handwritten document indexing and retrieval, and related problems.

In general, data collection can be designed or casual. In the first instance, writers with specific skills are recruited for contributing handwriting samples corresponding to specifically designed data consisting of symbols, characters, words or sentences or some combination of items in the target script. In the latter scenario, digital ink is a by-product of an ink application such as email, note-taking, form-filling etc. There is much less control on the nature or distribution of handwriting data collected, however, it happens naturally without the need for any specific effort on the part of the writer. This paper will focus on the task of designed data collection, taking the Tamil script as an example.

The paper is organized into several sections dealing with different aspects of the data collection problem. The second section describes the salient features of Indic scripts and their implications for online handwriting recognition. Corpus specification and related issues are discussed in the third section. This is followed by a discussion of the data collection methodology, data selection process, data collection and subsequent validation and annotation in sections four, five, six and seven, respectively. Some conclusions and directions for future work are presented in the final section.

## 2. STRUCTURE OF WRITING IN INDIC SCRIPTS

The 10 official Indic scripts - Devanagari, Tamil, Gurmukhi, Telugu, Kannada, Gujarati, Oriya, Bengali, Malayalam and Urdu - differ by varying degrees in their visual characteristics, but share some important similarities. With the exception of the Urdu script, they have evolved from a single source, the Brahmi script, first documented extensively in the edicts of Emperor Asoka of the third century BC. They are defined as "syllabic alphabets" in that the unit of encoding is a syllable, however, the corresponding graphic units show distinctive internal structure and a constituent

set of graphemes (Figure 1). The formative principles behind them may be summarized as follows [1]:

• graphemes for independent (initial) Vs
• C graphemes with inherent neutral vowel *a*
• V indication in non-initial position by means of *mātrās* (V diacritics)
• ligatures for C clusters
• muting of inherent V by means of a special diacritic called *virāmā*

**Fig. 1**. Diversity of Indic scripts

From the standpoint of HWR, an approach based on treating the syllabic units directly as pattern classes has to deal with their large numbers. Most of the Indic scripts have the order of 600 CV units and as many as 20,000 CCV ones in theory, although only a much smaller subset (especially of CCV units) is used in practice. The V diacritics and ligatures for C clusters are not standardized in some scripts. Since handwriting, in the online scenario, is captured as a sequence of pen strokes; the use of larger units also increases the variability in stroke order and hence the intra-class variability for the recognizer. In fact, even a single stroke can be written in various directions in such a scenario.

Approaches based on segmenting syllabic units into the constituent graphemes have to deal with the structural complexity of these syllabic units. In the online scenario, the beginnings of most graphemes are usually marked by pen-lifts. However, certain V diacritics may be fused inseparably with the underlying C grapheme. Different V diacritics may be visually similar and differ only in how they attach to the C grapheme. Similarly, many of the ligatures for C clusters are non transparent and have to be treated as separate graphemes.

In practice, the approach adopted for HWR is motivated more by pragmatic considerations such as the ease of segmentation of the handwritten word into a smaller number of graphically simpler sub-units, rather than by purely linguistic criteria, and linguistic interpretation of the recognized units is often relegated to a subsequent stage of processing. As a result, different researchers choose different sets of symbols as sub-word level units for recognition.

Ideally, datasets created to support handwriting recognition should accommodate different choices of symbol sets; however, it is not practical to accommodate these in a single annotation hierarchy. One solution is to support several sets of annotation each with its own hierarchy. These hierarchies would be common at the upper levels such as words and syllabic units and diverge thereafter to include different interpretations of symbols and where appropriate, individual strokes. This issue has been elaborated in fourth section.

## 3. CORPUS SPECIFICATIONS

Data collection is a resource-intensive activity and often only targeted at a particular use and purpose. The lack of standard corpus specifications makes the sharing of this data difficult. Further, these datasets cannot be re-used for other than the originally intended purpose. These shortcomings can be resolved by defining a corpus specification before the actual inception of data collection [2].

The corpus specification design requires the following issues to be studied:

- The script(s) to be collected: This is clearly a primary consideration in the design of the corpus that shapes many of the other decisions. Among other things, the encoding scheme for text (e.g. ISCII, UNICODE, specific fonts) needs to be decided. In addition to the rendering of text prompts during data collection, the same encoding is likely to be used as the basis for ground truth for subsequent annotation.

- The data-list: The design of the list of items to be collected constitutes one of the most important steps in the data collection exercise. Due to the variability of stroke-order and the whole gamut of writing styles to be captured, the data-list should be large enough to incorporate not only all of the symbols in the script but any important "co-articulation effects" and segmentation styles. However, a very large data-list is likely to tire and annoy writers. In general, it makes sense to use the minimum number of data items covering all the symbols and the most important variations. This reduces the writing time for each writer and makes data collection more effective.

- Number of trials: Collecting multiple instances of each data-item per writer is important for several reasons. First, writer-dependent recognition requires training and testing on samples of handwriting from the same writer. Second, the collection of multiple trials serve as an insurance against writer errors, and accidental loss of data. Again, a balance is needed between collection of additional trials and writing time and effort per writer.

- The data collection device: The choice of pen-input device is a critical one for collecting online handwriting data. There are a number of devices capable of

pen input on the market; some choices include external digitizing tablets (e.g. from Wacom) connected to desktops, PDAs such as Palm and PocketPC, Tablet-PCs and Anoto Digital Pen and Paper [3, 4]. The main considerations guiding the choice of pen-input device include cost, ease of use for the writer, sampling rate, spatial resolution, reliability, the need for additional channels such as pressure, API support for ink collection, among others. Some of this meta information also needs to be captured before a data-collection exercise. If data collection needs to be done in the field, battery life, portability, resistance to dust and general ruggedness also assume importance.

- UI design: The design of the user interface for data collection has to be specific to the device and the data to be collected. For example, the collection of word (or higher) level data would be much more difficult than isolated characters on a conventional PDA owing to its small screen. The user interface should minimally be able to display text prompts corresponding to the items to be collected, and present boxes or areas for writing. It should support means of erasing and rewriting samples if needed.

- Spacing of trials: It has been observed that collecting all trials for a given data-item contiguously from a writer seems to capture less variations in style, as compared to collecting them at different times and places.

- Recruiting of writers: The characteristics and distribution of the writers should be specified along with the ink-data. It is necessary that the writer characteristics are recorded as elaborately as possible. Though these details may not seem interesting at the time of recording, they play a very important role in distributing the datasets later. Moreover, a well documented handwriting corpus may also be used for other sociological purposes. In the order of importance, the writer characteristics that need to captured before any ink-data is collected are:

  - Nativeness of the script: The writing style of a native writer can be different from a writer who uses this script as a second language. A native writer follows a style of writing evolved from that taught in elementary schools, whereas the writing style of non-native writers may be derived from the structural pattern of the symbol or its similarity with a symbol in his/her own native language. This typically gives rise to intra- and inter- stroke order differences which finally leads to greater variability. Many corpora imply native writers of a certain script, but it is always recommended to specify the maximum percentage of non-native writers.
  - Frequency of use of the script: Fluent writing like any other acquired skill requires continuous practice. The handwriting of a person who uses a script often will show different traits as compared to a person who uses it occasionally.
  - Education: It is useful to include writers with different levels of education, since this is clearly correlated with handwriting skills. This is also valuable from an application standpoint - processing a college application form would imply a different educational background for writers as compared to a railway reservation form which might need handling greater variation in education-level.
  - Profession: Different professions involve use of a particular script in different environments. For example, a secretary taking dictation in an office has to write at the speed of speech. On the other hand, an author or a teacher might adopt different styles. Their habitual writing modes lead to different writing styles even when both of them are placed in the same environment.
  - Distribution of sex: Equal percentages of male and female writers are generally recommended, but almost impossible in practice. The corpus specification should set a threshold on any deviation from the optimum percentages.
  - Distribution of age: For a general purpose corpus of adult handwriting, the writers' ages should lie between 16 to 60, to capture the variations in handwriting with age.
  - Right/Left Handness: This is an attribute that is unique to the handwriting domain. A distribution of left and right handed writers that is representative of the general population is recommended.
  - Region: Writers from different parts of the world can have different styles of handwriting for the same script. The shapes of certain characters may vary significantly as a result of local influences. In the Indic context, scripts such as Devanagari and Tamil are used in diverse geographical regions within and outside India. Devanagari is used to support multiple languages such as Nepali and Marathi in addition to Hindi, whereas Tamil is an official language of countries such as Sri Lanka, Singapore and Malaysia.
  - Skill with the device: Proficiency with the device allows the user to write in his/her natural

style. Writing on the smooth glass surface of a PDA feels very different from writing on paper. In practice, it is difficult to find writers with different levels of proficiency with the device but nevertheless, this is an important attribute to capture since it can have a significant impact on the writing style.

## 4. DATA COLLECTION METHODOLOGY

In this section we describe the methodology that we have adopted for the collection of handwriting data in the Tamil script. The general flow of handwriting data collection (and subsequent annotation) is shown in Figure 2.

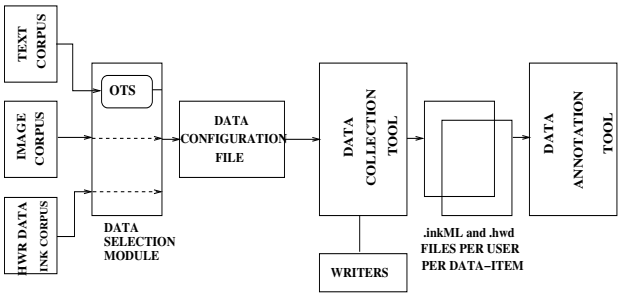


**Fig. 2**. General Flow of Handwriting Data Collection

The very first step in data collection is the creation of the data-list – the list of items of handwriting to be collected from each writer [5]. This is accomplished in the data selection module. The data-list is then input to the data collection tool - a software application running on a TabletPC. The output of data collection is a partially annotated dataset of handwriting samples, and forms the input to the annotation phase. These steps are now described in detail in the context of both isolated characters and isolated words for the Tamil script.

## 5. DATA SELECTION

None of the Indic scripts are traditionally written in boxes. However, writing in boxes provides valuable segmentation cues to the recognition engine and thus being able to "box" the script simplifies the segmentation of writing and leads to improved accuracy. Tamil uses the vowel-muting diacritic or "halant" to unravel conjuncts and is written as a linear sequence of "characters". An informal observation of several Tamil writers has revealed that they could write Tamil characters in boxes with minimal training. The first part of data collection for the Tamil script therefore involves the collection of isolated Tamil characters.

In order to create the Tamil isolated character-set, we compiled a list of independent V and C graphemes, CV combinations where the vowel diacritics attach above or below the base C grapheme or are otherwise difficult to segment, and those vowel diacritics that occur as distinct characters to the left or right of the base C. The set also includes selected C cluster ligatures and their CV combinations, for a total of 156 characters.Figure 3 shows the symbols for both isolated characters and word level Tamil data.

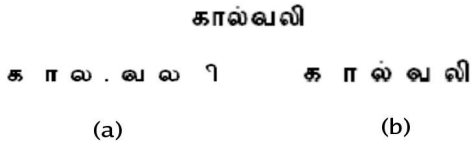


**Fig. 3**. Symbols for (a) isolated Tamil characters (b) Tamil words

The second part of data collection for Tamil is focused on isolated words. The word forms the fundamental unit of writing for any script. From a recognition perspective, the word is especially important in Indic scripts such as Devanagari given the presence of conjuncts of varying size and complexity, and the absence of a tradition of writing words in boxes. Thus, the ability to recognize words written continuously (i.e., without boxes) is important for even the most constrained applications such as form-filling. In general, the isolation of words from larger units of writing such as sentences and lines is based on spatial and temporal cues. Our focus here is to collect handwriting data to support research into the recognition of words once they have been isolated by some other means.

The data selection procedure for words involved:

1. Identification of the symbol-set (sub-word units): Due to the absence of explicit segmentation provided by the boxes, a good set of symbols (sub-word units) is one that balances the ease of segmentation of the word into those sub-units with a stable pattern across writers. This in turn affects the accuracy of the recognition of the symbols. Since Indic scripts do not have a prominent cursive style and penlifts may be expected between graphemes, we adopted as the symbol set, the basic graphemes in the Tamil script (independent Vs, Cs, V diacritics, vowel-muting diacritic) and added some symbols corresponding to CVs which could not be easily segmented into the constituent base C and V diacritic.

2. Generation of word-list given the symbol-set: As discussed earlier, our attempt is to cover the entire symbol-set using the smallest number of words. In order to achieve this, we used the TDIL Tamil text corpus [6]. Unique words along with their respective frequencies were extracted from the corpus using the language modeling toolkit from CMU [7]. Words with frequency less than a threshold were purged from the result in an attempt to discard very rare as well as erroneously spelt words in the corpus. Next, a set cover

algorithm was used to extract from the remaining words, a minimal subset of words that covered all of the symbols to be recognized [8, 9]. The resulting list of words was verified and minor additions and substitutions carried out manually in order to obtain a final data-list of 47 words covering the selected set of Tamil symbols.

## 6. DATA COLLECTION

For both isolated characters and words, handwriting data was collected from writers using a software tool running on a TabletPC (HP TabletPC TC1100). The TabletPC is a convenient device for data collection for several reasons. It has sufficient processing power to support simultaneous inking and sampling of handwriting at a high sampling rate of 120 Hz. It provides spatial resolutions of more than 400dpi in both X and Y directions. It features an active digitizer which is not affected by inadvertent contact of the hand on the writing surface which leads to spurious ink on normal touchscreens. Its larger writing surface is also much more natural to use compared to PDAs. The device OS comes with good support for the creation of pen-based tools and UI controls for the collection of digital ink. However, the feel of writing on glass is different from writing on paper and the writer takes some time getting used to it.

### 6.1. Setup

The tool is driven by a configuration file that specifies the data-list for collection along with other configuration details as: (1) Unique ID of the configuration file, along with creation date and author information. (2) A description of the data to be collected and the data collection methodology. (3) The list of items to be collected along with information about the (i) the corresponding script, and (ii) type of text prompts provided. The prompts may be specified as text (data-list items), bitmaps or digital ink (handwritten prompts). (4) Number of data collection trials per user.

Once configured as described above, the data collection tool is ready for use. The tool initially presents a screen for the collection of writer-specific information such as Name, Age, Gender, Hand R/L, Region, nativeness of script, proficiency with device, profession, usage of script

It is assumed that data collection is supervised by someone familiar with English as well as with the device. The supervisor is responsible for assigning each contributing writer a unique numeric ID, collecting and entering writer information into the tool, and providing the writer assistance with the user interface or device as needed. No familiarity is assumed on the part of the writer with the device, UI controls, or the English language.

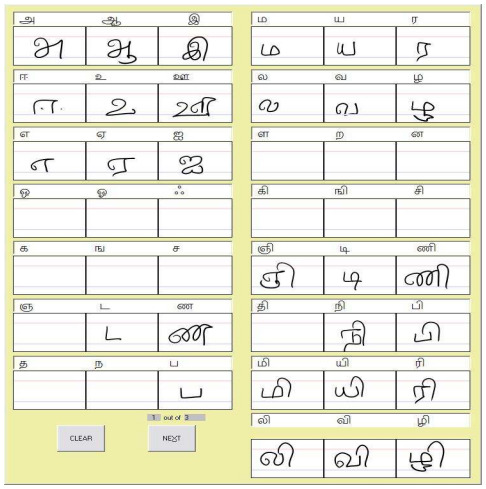


**Fig. 4**. Data Collection using the tool

### 6.2. Data Collection Trials

Once the meta-data has been collected, the actual data-collection begins. Data collection is organized into a series of trials. The writer can stop after any trial and resume later. Before the first trial, every writer is provided with a writing area where he/she can practice writing and set the reference lines as per his or her writing style. These reference lines bind the middle zone of handwriting and mirror the lines provided in copybooks for the practice of writing in school.

Within each trial, the items from the data-list are presented in order as a sequence of text prompts and writing areas, on one or more screens (pages) as needed (Figure 4). The text prompts are rendered just above the respective writing areas using a script-specific font if specified as text, or as bitmaps or digital ink as specified. The reference lines as set by the writer are displayed in each writing area. The user is at liberty to address the items presented in any order. The user interface allows the writer to clear and rewrite a particular item. This gives the writer greater flexibility in completing the task and helps to reduce the possibility of erroneous data.

### 6.3. Output

The output of the Data Collection Tool is in the form of ASCII files organized by script and writer. A separate file is generated for each sample of each item in the data-list, and the item (character or word) ID and trial number are encoded in the file name.

Each file captures the following (optional) meta-data in a header: (a) annotation such as word-level ground truth (b) positions of reference lines (c) type and spatial resolution of digitizer (d) coordinates of box (for isolated characters)

Following the header, the handwriting data is captured

as a sequence of (x,y) points and pen up/down events. Along with each pen down or pen up, a timestamp is stored denoting the number of milliseconds from the start of the trial. In addition to the ink files, meta-data such as writer profiles are stored in common files at the root of the directory.

Going forward, we intend to use InkML for the capture of handwriting, along with *hwDataset* [10] documents for meta-data. InkML [11] is a draft specification for the representation of digital ink from the World Wide Web Consortium (W3C) and provides an open, platform-independent XML representation for the description of digitizer characteristics, ink channels such as X and Y, and the digital ink itself. *hwDataset* is a set of XML tags for the annotation of handwriting including writer information, ground truth etc.

## 7. VALIDATION AND ANNOTATION

Validation in this context refers to visual review of the data collected by the supervisor immediately after the conclusion of the trials for a particular writer. This is intended to catch human as well as system errors in the data collection process - items that may have been skipped or wrongly entered, accidents such as corruption or overwriting of data. It is important to validate while the writer is still accessible and it is possible to get the errors rectified. The validation activity is supported by a separate tool for viewing the data collected.

Annotation refers to the labeling of the collected handwriting data in accordance with a designated annotation hierarchy. The files of digital ink and any meta-data captured as a part of the data collection phase, form the input to the annotation phase. The chief activity in this phase is the tagging of ink with labels corresponding to ground truth, writing style etc. at different levels of an appropriate hierarchy of annotation [10]. Labels may be generated by human annotators or by machine algorithms; in practice, a combination may be used to completely annotate large corpora.

In our design, the corpus or dataset is represented as a collection of *hwDataset* documents organized into an appropriate directory structure. Each *hwDataset* document is paired with an InkML document containing the digital ink data referred to in the document. A detailed discussion of annotation is beyond the scope of this paper; however, a brief overview of the *hwDataset* representation for annotation and the annotation tool that we have created is presented in [10].

The final output of the annotation phase is sometimes followed by independent validation of the annotation by a neutral third party, following which the corpus or linguistic resource is generally considered to be ready for release.

## 8. CONCLUSIONS AND FUTURE DIRECTIONS

The collection of handwriting samples from native writers is the first step in the creation of linguistic resources that can be used to further research in handwriting recognition, analysis, script identification and other language technologies for the Indic scripts. In this paper, we have described some of the issues to be considered during the design of handwriting data collection, and our early attempts at addressing these issues, primarily in the context of the Tamil script. Over the coming months, we intend to scale our existing efforts to a larger number of writers, as well as extend our data collection efforts to other scripts such as Devanagari and Telugu.

We also intend to explore modifications to the current data collection methodology. For instance, rather than collecting lists of isolated characters or words, an "omnibus" approach that combines different units of writing in a single trial may be preferable. Second, the influence of the writing style displayed in the text prompts on the writing of the user needs to be studied and minimized. We also plan to adopt stricter criteria on the recruitment of writers to ensure the desired distribution in terms of age, handedness, gender and other parameters. In tandem, we hope to refine our tools for data collection and annotation, and the representations of ink and annotation, and develop libraries for convenient access to the collected data.

## 9. REFERENCES


[1] Florian Coulmas, *The Encyclopedia of Writing Systems*, pp. 229–230, Blackwell Publishing, 1st edition, 1999.

[2] Florian Schiel and Christoph Draxler, *Production and Validation of Speech Copora*, 2.4 edition, 2003.

[3] *Anoto Functionality*, http://www.anotofunctionality.com/.

[4] Hewlett-Packard Corp., *HP Forms Automation Systems (FAS)*, http://www.hp.com/go/fas, 2003.

[5] Ajay S. Bhaskarabhatla and Sriganesh Madhvanath, *Experiences in Collection of Handwriting Data for Online Handwriting Recognition in Indic Scripts*, LREC, May, 2004.

[6] *Technology Development for Indian Languages*, http://tdil.mit.gov.in/corpora/ach-corpora.htm.

[7] P.R. Clarkson and R. Rosenfeld, *Statistical Language Modeling Using the CMU-Cambridge Toolkit*, pp. 2707–2710, Proc. EUROSPEECH, vol. 1, Sep. 1997.

[8] A. Caprara, M. Fischetti, and P. Toth, *Algorithms for the Set Covering Problem*, Tech. Rep. No. OR-98-3, DEIS-Operations Research Group, 1998.

[9] J.P.H. van Santen and A Buchsbaum, *Methods for optimal text selection*, pp. 553–556, Proc. EUROSPEECH, 1997.

[10] Ajay S. Bhaskarabhatla, M.N.S.S.K. Pavan Kumar, A. Balasubramanian, C.V. Jawahar, and Sriganesh Madhvanath, *Representation And Annotation Of Online Handwritten Data*, IWFHR-9, Japan, Oct, 2004.

[11] *Ink Markup Language*, http://www.w3.org/2002/mmi/ink.