

Basically, we have five datasets available with us

1. Bentham manuscripts
2. IAM data set
3. Rimes
4. Saint Gall
5. Washington

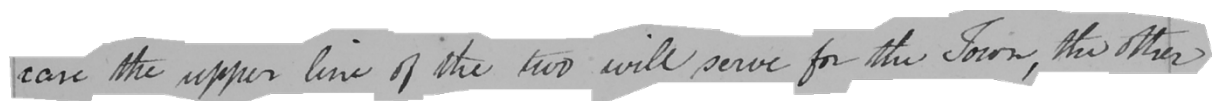
### Bentham manuscripts:

This contains large set of document written by renowned English philosopher and reformer Jermy Bentham (1748 – 1832) The transcription of this collection is currently being carried out by amateur volunteers participating in the award-winning crowd-sourced initiative, Transcribe Bentham. Currently, more than 6,000 documents have been transcribed using this public web platform.

This dataset is free available for research purposes and it is provided into two parts: the images and the GT. The GT includes information about the layout and the transcription at line level of each image in PAGE format. Both parts must be downloaded separately. A detailed description is included in each part explaining how the dataset is organized.

**Total Examples -11,473 documents**

e.g



case the upper line of the two will serve for the Town , the other

### IAM dataset:

The IAM Handwriting Database contains forms of handwritten English text which can be used to train and test handwritten text recognizers and to perform writer identification and verification experiments. The database contains forms of unconstrained handwritten text, which were scanned at a resolution of 300dpi and saved as PNG images with 256 gray levels. The figure below provides samples of a complete form, a text line and some extracted words.

**Total Examples :13353**



GT:c03-000a-00 err 186 38 353 757 1810 123 The|film|version|of|Miss|Shelagh|Delaney's



GT:a01-000u-00 ok 154 19 408 746 1661 89 A|MOVE|to|stop|Mr.|Gaitskell|from

All forms and also all extracted text lines, words and sentences are available for download as PNG files, with corresponding XML meta-information included into the image files. All texts in the IAM database are built using sentences provided by the LOB Corpus.

---

## Characteristics

The IAM Handwriting Database 3.0 is structured as follows:

- 657 writers contributed samples of their handwriting
- 1'539 pages of scanned text
- 5'685 isolated and labeled sentences
- 13'353 isolated and labeled text lines
- 115'320 isolated and labeled words

## Rimes :

These people ask for the consent that this will be used for research and development purposes only, I have raised the request they may come up with further details

## Saint Gall Database:

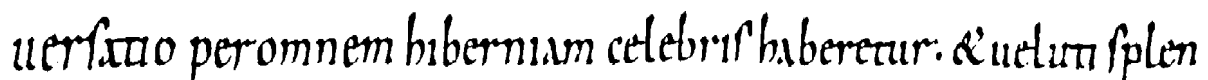
The Saint Gall database presented in [1] contains a handwritten historical manuscript with following characteristics:

- 9th century
- Latin language
- single writer
- Carolingian script
- ink on parchment

The original manuscript is housed at the [Abbey Library of Saint Gall](#), Switzerland. The manuscript images [3] were made available online by the e-codices project and a text edition [4] was attached at page-level by the Monumenta project. We have additionally added our binarized and normalized text line images to the manuscript data. Altogether, the manuscript data is given by:

**Total Examples : 1410**

Lines:.png



GT:csg562-003-01 v-e-r-s-a-t-i-o|p-e-r|o-m-n-e-m|h-i-b-e-r-n-i-a-m|c-e-l-e-b-r-i-s|h-a-b-  
e-r-e-t-u-r|e-t|v-e-l-u-t-i|s-p-l-e-n  
conversatio|per|omnem|hiberniam|celebris|haberetur|et|veluti|BREAK

### Washington Database:

The Washington database was created from the George Washington Papers at the Library of Congress and has the following characteristics:

- 18th century
- English language
- two writers
- longhand script
- ink on paper

The word ID "270-01-01" can be read as follows: page 270 (Library of Congress, George Washington Papers, Series 2, Letterbook 1), line 1, word 1. The line ID is "270-01" and the page ID is "270" accordingly. Word and line numbers start at 1.

**Total Examples :656**

270. Letters, Orders and Instructions. October 1755.

270-01 s\_2-s\_7-s\_0-s\_pt|L-e-t-t-e-r-s-s\_cm|O-r-d-e-r-s|a-n-d|I-n-s-t-r-u-c-t-i-o-n-s-s\_pt|O-c-t-o-b-e-r|s\_1-s\_7-s\_5-s\_5-s\_pt

**Conclusion :Total we have  
656+1410+13353+11473 = 26892 examples  
with us to use**