IJDAR (2002) 5: 39–46

International Journal on IJDAR Document Analysis and Recognition

The IAM-database: an English sentence database for offline handwriting recognition

U.-V. Marti, H. Bunke

Department of Computer Science, University of Bern, Neubrückstrasse 10, 3011 Bern, Switzerland; e-mail: {marti,bunke}@iam.unibe.ch

Received September 28, 2001 / Revised October 10, 2001

Abstract. In this paper we describe a database that consists of handwritten English sentences. It is based on the Lancaster-Oslo/Bergen (LOB) corpus. This corpus is a collection of texts that comprise about one million word instances. The database includes 1,066 forms produced by approximately 400 different writers. A total of 82,227 word instances out of a vocabulary of 10,841 words occur in the collection. The database consists of full English sentences. It can serve as a basis for a variety of handwriting recognition tasks. However, it is expected that the database would be particularly useful for recognition tasks where linguistic knowledge beyond the lexicon level is used, because this knowledge can be automatically derived from the underlying corpus. The database also includes a few image-processing procedures for extracting the handwritten text from the forms and the segmentation of the text into lines and words.

Keywords: Handwriting recognition – Database – Unconstrained English sentences – Corpus – Linguistic knowledge

1 Introduction

The availability of large amounts of data for training and testing is a fundamental prerequisite for building handwriting recognition systems with high recognition accuracy. Furthermore, the comparison of various recognition methods has become a focus of attention recently. Therefore, the acquisition and distribution of standard databases has become an important issue in the handwriting recognition research community [4]. Examples of widely used databases are CEDAR [6], NIST [25], and CENPARMI [23] in the offline domain; UNIPEN [5] for online; and ETL9(Japan) [20] as well as PE92(Korea) [9] in the field of oriental handwriting recognition. However, these databases contain mostly isolated characters or single words. There is definitely a lack of databases that contain large amounts of general unconstrained handwritten English sentences, produced by different writers. The NIST database contains instances of the Preamble of the American Constitution, but the vocabulary is too small to be useful for realistic applications. Another database containing English sentences is described in [22]. However, all sentences in this database were written by the same person and the underlying vocabulary is rather small.

In this paper we describe the actual version (as of October 2000) of the IAM-database[1] that contains full English sentences. The database consists of 82,227 instances of handwritten words distributed over 9,285 lines of text produced by approximately 400 writers. The underlying lexicon includes 10,841 different words. The database can be used to train and test word and sentence recognizers. All texts come from the LOB corpus, which is electronically available. Hence, it is possible to automatically generate various kinds of language models. This property makes the database particularly interesting for the development of handwriting recognition systems that use linguistic knowledge beyond the lexicon level [2,10]. However, the handwritten texts can also be used for other tasks in the domain of unconstrained handwriting recognition, such as segmentation [8] or writer identification [18]. The database described in this paper is freely available to other researchers upon request.

An earlier version of the database was described in [15]. Meanwhile, the database has grown and became almost twice as large. Because of space limitations, the description given in [15] is rather short. In the present paper a much more detailed description is provided. In particular, a complete description of the image processing routines that are part of the database is included. These image processing routines allow a user of the database to perform recognition experiments without the need of developing his or her own low-level image processing and segmentation procedures. A novel feature of the database compared to the version described in [15] is a set of routines that segment lines of text into individual words.

[1] IAM = Institut für Informatik und angewandte Mathematik (= Department of Computer Science and Applied Mathematics), University of Bern, Bern, Switzerland

Table 1. The categories of text collected in the LOB corpus

A	Press: reportage	44
B	Press: editorial	27
C	Press: reviews	17
D	Religion	17
E	Skills, trades, and hobbies	38
F	Popular lore	44
G	Belles letters, biography, essays	77
H	Miscellaneous	30
J	Learned and scientific writings	80
K	General fiction	29
L	Mystery and detective fiction	24
M	Science fiction	6
N	Adventure and western fiction	29
P	Romance and love story	29
R	Humour	9
	Total	500

This makes it possible to use the database for the development of classifiers for isolated word recognition. The entities at the lowest level in the previous version of the database are whole lines of text. Consequently, this version of the database is useful only for segmentation-free recognizers that perform segmentation of a line of text into words and word recognition in an integrated fashion. An example of this kind of recognizer is the hidden Markov model. More generally, the present version of the database supports both segmentation-free and segmentation-based word recognition procedures.

In the next section, we describe the fundamental steps of the database acquisition procedure. Section 3 is concerned with the task of defining the ground truth values. Section 4 describes a technique to extract the handwriting from the forms and the segmentation of the texts into lines and words. In Sect. 5 further characteristics of the data collection are listed. Finally some conclusions are drawn in Sect. 6.

2 Corpus and forms

In the domain of linguistics, several large collections of texts, called corpora, are available. These corpora have different appearance and content. In some only plain text is included. In more elaborated versions the words are tagged. That means for every word in the text there is a tag, which marks the word as a noun, a verb, or another word class. Often the tempus, modus, and so on, of a word is also included.

It was decided to use the Lancaster - Oslo/Bergen corpus (LOB) [7], a collection of 500 English texts, each consisting of about 2,000 words as a basis of our database. The LOB corpus is the British pendant to the Brown corpus [3], which has a structure similar to that of the LOB corpus. The texts in the LOB corpus are of quite diverse nature. The different categories and the number of texts per class are listed in Table 1. Using a corpus as the foundation of the database rather than collecting text from "random" sources has the advantage that linguistic knowledge can be automatically extracted in a more systematic and easy way.

It was our goal to acquire a database of handwritten sentences that are all contained in the LOB corpus. For this purpose, we split the texts in the corpus into fragments of about three to six sentences with at least 50 words each. These text fragments were printed onto forms and we asked a number of persons to write the text printed on the forms by hand.

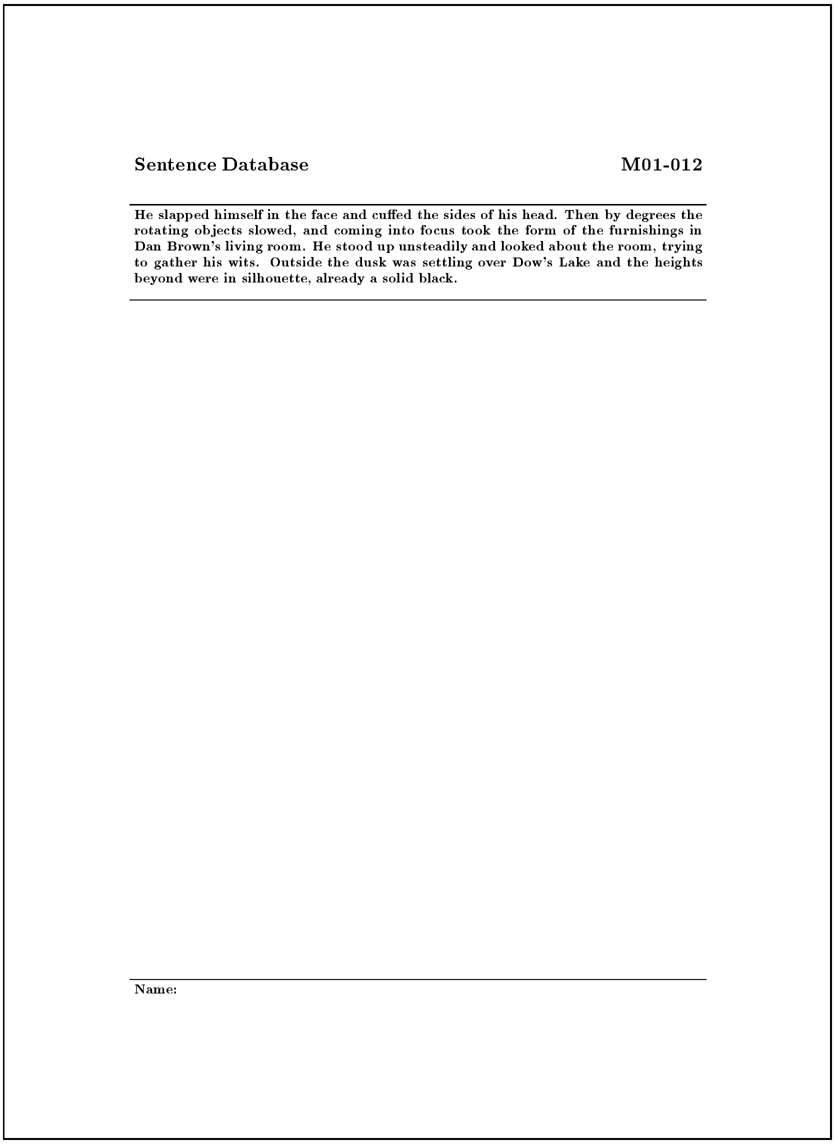
The forms were automatically generated. We extracted the sentences of each text fragment from the corpus and generated a LaTeX document containing the text and the structure of the form. The formatted documents were printed by a HP Laserjet 4000TN at a resolution of 600 dpi.

Each form consists of four parts (see Figs. 1 and 2). The first part comprises the title "Sentence Database" and a number assigned to the text. The first character of this number shows which category the text belongs to, and the following two digits identify the text number. For example in Fig. 1, M01 indicates that the text on the form is extracted from text "01" in the text category "Science fiction". The next three digits show with which sentence the text starts. In the second part of the form, the text the individual persons were asked to write is printed. The third part of the form is a blank zone where the writers have to put in their handwriting. In the last part, the writer can voluntarily enter his or her name. All four parts are separated from each other by a horizontal line. This makes it easy to automatically extract the individual parts from a form (see Sect. 4).

As the main focus of the research that led to the acquisition of the database described in this paper is on high-level recognition using language models, we wanted to make the image processing part as easy as possible. Therefore, it was decided that the writers had to use rulers. These guiding lines, with 1.5 cm space between them, were printed on a separate sheet of paper which was put under the form. The writers were asked to use their every day writing in order to get the most natural and unconstrained way of writing. We also told the writers to stop writing, if there was not enough space left on the form to write the whole text. This way we wanted to avoid getting compressed and deformed words. No restrictions were imposed on the writing instrument. Hence, text produced with a number of different writing instruments is included in the database (ballpoint pens, ink pens, and pencils, all with various stroke widths).

The filled forms were scanned with a HP-Scanjet 6100 connected to a Sun Ultra 1. The software used to scan the data is *xvscan* version 1.6[2]. It is an add-on to the well-known image tool *xv*. The resolution was set to 300 dpi at a grey-level resolution of 8 bits. The images were saved in TIFF-format with LZW compression. Each form was completely scanned, including the printed and handwritten text. Thus, it is also possible to do experiments with the machine-printed text, for example, to distinguish be-

[2] Further information can be found at http://www.tummy.com/xvscan/



Sentence Database **M01-012**

He slapped himself in the face and cuffed the sides of his head. Then by degrees the rotating objects slowed, and coming into focus took the form of the furnishings in Dan Brown's living room. He stood up unsteadily and looked about the room, trying to gather his wits. Outside the dusk was settling over Dow's Lake and the heights beyond were in silhouette, already a solid black.

Name:

Fig. 1. Empty form

tween handwritten and machine-printed text, or to apply image processing and segmentation operations on the forms, different from these described in Sect. 4.

3 Labeling

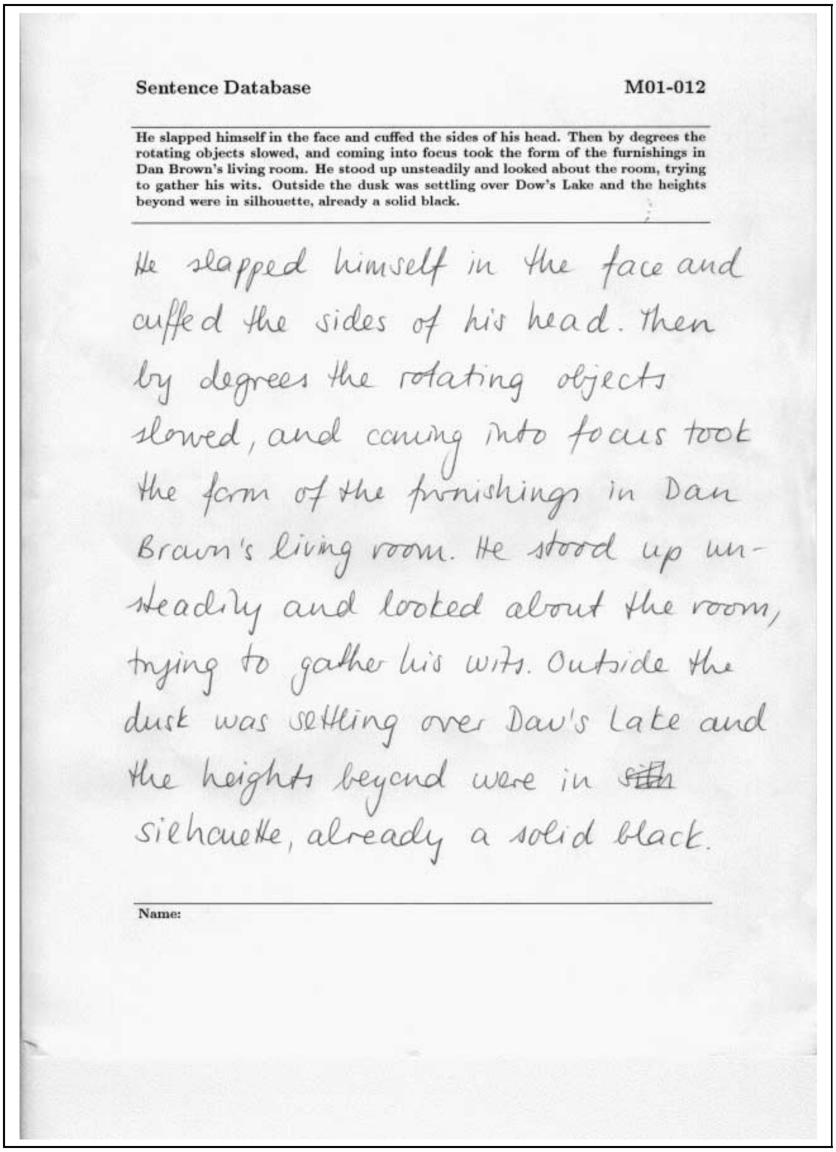
Labeling of data is a prerequisite for any recognition experiment. Because this task is time consuming and error prone, it was decided to do as much as possible automatically. The sources of all the forms printed (and subsequently filled by the writers) were saved on disk. Thus, it was an easy task to generate the correct labels for the printed text on the forms.

To create the label files, the text of a form was copied twice into the label file, once for the machine-printed text and once for the handwritten text. Then the line feeds were filled in manually. In some cases corrections were necessary, because the handwritten text did not exactly correspond to the printed text. These corrections include deletions, insertions, and changes of words so as to make the text in the label file of each form identical to the handwritten text. All corrections were done manually, but they did not take a long. Typically, only approximately 30 s of manual processing time for the generation of the label files and the correction of errors were spent on each document, which is less than the time required for scanning. An example of a final label file is shown in Fig. 3.

Figure 4 gives a synoptic overview of the different steps involved in the generation of the database. It is clear that the corpus used here can be replaced by any other collection of texts.

4 Text extraction and segmentation

For extracting the handwritten text from a scanned form, a number of image preprocessing and segmentation algorithms have been developed. It is not mandatory to apply these algorithms to the images in the database. However,



Sentence Database **M01-012**

He slapped himself in the face and cuffed the sides of his head. Then by degrees the rotating objects slowed, and coming into focus took the form of the furnishings in Dan Brown's living room. He stood up unsteadily and looked about the room, trying to gather his wits. Outside the dusk was settling over Dow's Lake and the heights beyond were in silhouette, already a solid black.

He slapped himself in the face and
cuffed the sides of his head. Then
by degrees the rotating objects
slowed, and coming into focus toot
the form of the frnishings in Dan
Brawn's living room. He stood up un-
steadily and looked about the room,
trying to gather his wits. Outside the
dusk was setting over Daw's Late and
the heights beyond were in ~~sith~~
silhouette, already a solid black.

Name:

Fig. 2. Filled form

they are a convenient tool for experiments where the focus is on recognition rather than image processing aspects. First, the skew of the document is corrected. Then the positions of the three horizontal lines are computed by a projection method. Given this positional information the handwriting is extracted. Next, the handwritten text is segmented into text lines. Finally, each text line is segmented into individual words.

To find the first horizontal line, the form is scanned top-down in the middle of the image. The first black pixel found is assumed to belong to the first horizontal line. By following the line to the left and to the right, the end points of the line are found. Once the left and right end of the first horizontal line has been determined, its angle is used to correct the skew of the whole document by a rotation.

Then, the form is segmented into its four main parts (see Sect. 2). This is a relatively easy task because the four parts are separated from each other by very long horizontal lines, which are easy to detect by horizontal projection. To make the projection algorithm more robust, not only the horizontal projection profile is considered, but also the value of the longest horizontal black run in each row. A horizontal line separating two parts from each other is characterized by a value in the horizontal grey-value projection histogram greater than a threshold t and a value of the longest horizontal black run greater than a threshold t' (see Fig. 5). After all three horizontal lines have been found, we are able to extract the part of the form that contains the handwriting. Over the whole database the handwritten zone was automatically extracted by means of this procedure without any error.

The next step in preprocessing is to cut the text into individual lines. For this purpose a histogram of the horizontal black/white transitions is used. In this histogram we look for local minima. If the value at a local minimum is zero, a cut has been found that does not touch any word. If the value is greater than zero, we have found a position where we can horizontally cut the image with a)

He slapped himself in the face and cuffed the sides of his head. Then by degrees the
rotating objects slowed, and coming into focus took the form of the furnishings in
Dan Brown's living room. He stood up unsteadily and looked about the room, trying
to gather his wits. Outside the dusk was settling over Dow's Lake and the heights
beyond were in silhouette, already a solid black.

b)

He slapped himself in the face and
cuffed the sides of his head. Then
by degrees the rotating objects
slowed, and coming into focus took
the form of the furnishings in Dan
Brown's living room. He stood up un-
steadily and looked about the room,
trying to gather his wits. Outside the
dusk was settling over Dow's Lake and
the heights beyond were in
silhouette, already a solid black.

Fig. 3. Example of the label file corresponding to Fig. 2; **a** ground truth of the machine printed text; **b** ground truth of the handwritten text

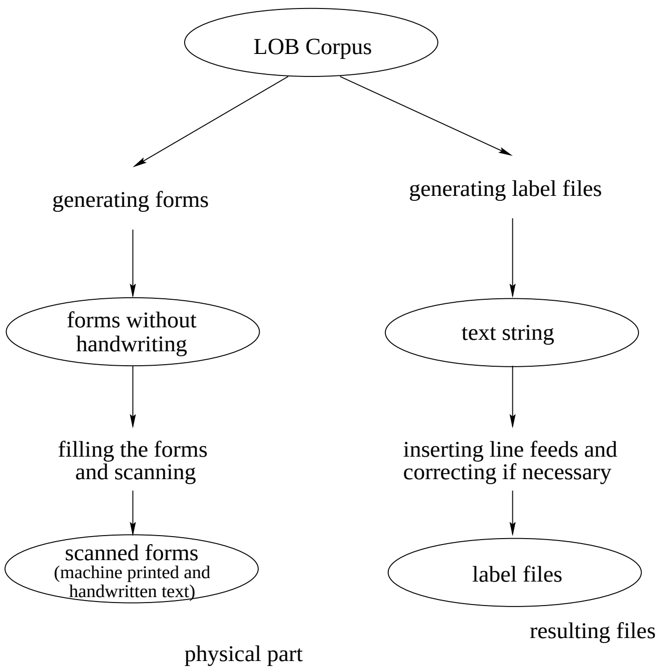


Fig. 4. Database acquisition

a minimal number of intersections with strokes belonging to words of the previous or the following text line. To handle intersections of this kind a method based on the center of gravity is used. If the center of gravity of the connected component that is cut is in the range of the previous (the following) text line, the connected component is assumed to belong to that text line. If the center of gravity is near the cutting line, the component is cut into two parts, one belonging to the previous and one to the following text line. An example is shown in Fig. 6.

With this method we could extract almost all text lines correctly. Only in about 1.2% (113 of 9157) of the lines did errors occur. The 113 segmentation errors can be classified into acceptable and serious errors. An acceptable error is defined as one where only single punctuation marks or i-dots are assigned to the wrong line, but no cut component. By contrast, parts of letters or words assigned to the wrong text line are considered serious errors. There were 52 acceptable and 61 serious errors out of a total of 113. The correction of these errors is left to the user of the database. That is, no attempts were made to include any manual corrections of the segmentation errors in the database.

To segment the text lines into single words, a method similar to those described in [21] and [11] is used. Because a word can be split into several components, the goal is to cluster the connected components of a text line image into words. First, the convex hull and the center of gravity of each connected component in a line of text is computed. Then for each pair of connected components, c_1 and c_2, the straight line segment s that connects the center of gravity of c_1 with the center of gravity of c_2 is considered. The distance d between the two points where s intersects the convex hull of c_1 and c_2 is determined. This distance d is assigned to s as a weight. A graphical illustration is shown in Fig. 7. By means of this procedure, a completely connected and weighted graph is obtained, where each node corresponds to a connected component in the image and the weight on an edge represents the distance between two connected components. Given such a graph, its minimum spanning tree is computed. Finally, all distances that occur as weights on the edges of the minimum spanning tree are clustered into two groups, namely, intra-word and inter-word distance. For this purpose, Otsu's thresholding algorithm [19] is used, which yields a threshold t. Any two connected components, c_1 and c_2, that are linked by an edge in the minimum spanning tree with a weight $d \leq t$ are considered to be part of the same word. By contrast, if the weight d is larger than t, c_1 and c_2 are regarded as belonging to two different words. Applied on 541 text lines with 3,899 words, a correct word extraction of 94.92%, with 3.62% of the words over- and 1.46% undersegmented, was achieved.

The primary goal of the preprocessing and segmentation procedures described in this section is to support the labeling of the text. However, these procedures can be integrated in any recognizer as well. All procedures have been implemented in C++ and can be compiled under Unix/Solaris and Linux. To actually run the procedures only the TIFF library for image read and write, but no other software utilities are needed.

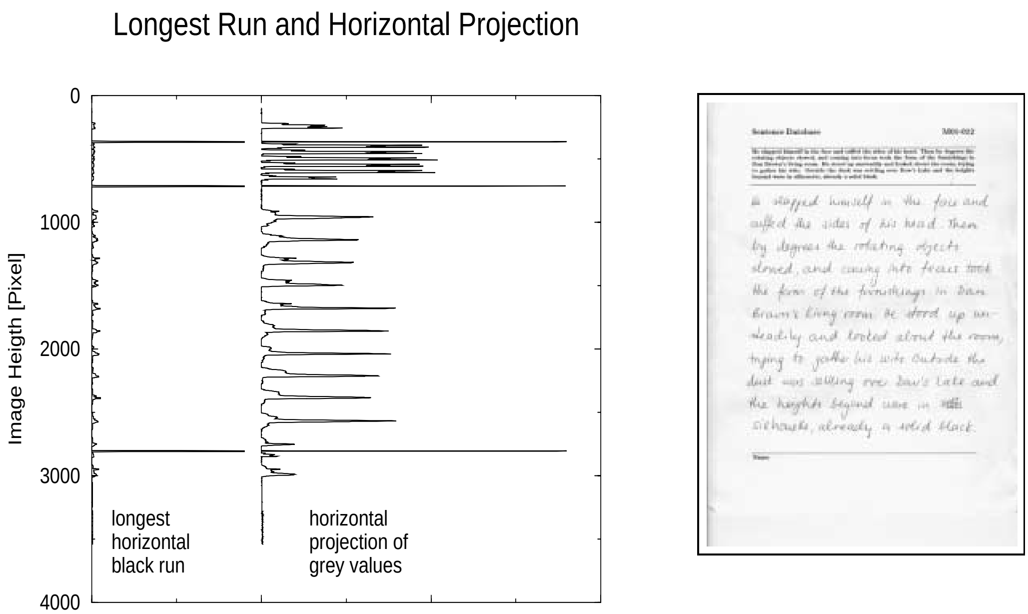


Fig. 5. Length of longest horizontal black run and horizontal projection of the grey values for the form shown in Fig. 2

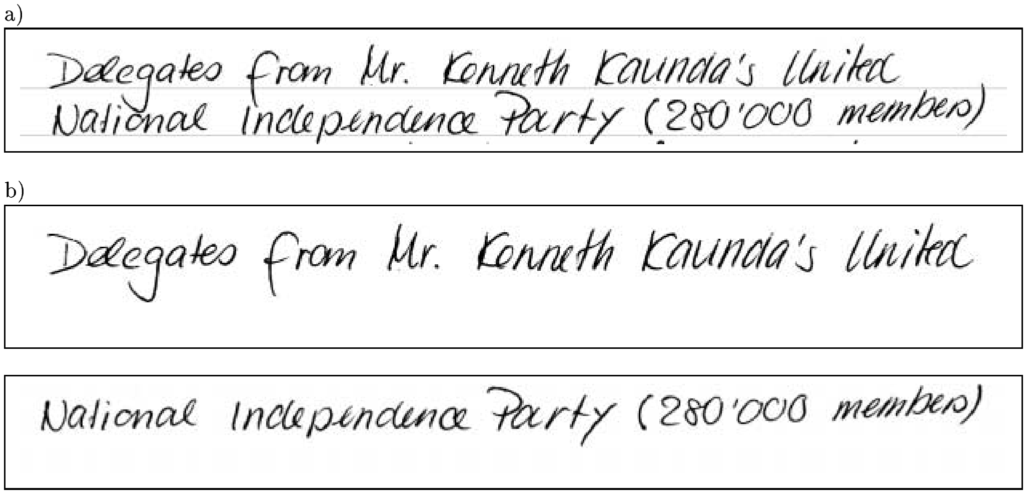


Fig. 6. a Two lines of text that can be separated without a cut. **b** Result of the line segmentation (see text)

5 Further characteristics of the database

As of October 2000 the database consists of a total of 1,066 filled forms. The distribution over the different text categories and over the text fragments can be seen in Table 2. For example, the LOB corpus includes 44 texts belonging to category A (see Table 1). These are numbered from 1 to 44. There are 70 forms included in the database, each containing a part of Text 1, 16 forms each containing a part of Text 2 and so on. Texts 7 through 44 of category A are not represented in the database. Note that the three digit number in the right upper corner of each form is the number of the first sentence on the form in the corresponding text (see Fig. 1). From the 9,285 lines of handwritten text all together, 9,157 lines can be used for text recognition. The remaining 128 lines contain information not contained in the LOB corpus, for example the name of the writer or additional comments. A total of 82,227 word instances out of a vocabulary of 10,841 words are included in the data base.

There is an average of 8.59 handwritten text lines per form. The distribution of the number of handwritten lines per form can be seen in Fig. 8. In most cases the space available to copy the machine printed text by hand was sufficient. The average number of handwritten words per text line is 8.98 or 77.14 per text. The distribution of the number of handwritten words per text line over the database is shown in Fig. 9.

There are some subsets of forms in the database that include multiple instances of the same text fragment written by different writers (subsets x and u), or the same text fragment written by the same writer (subsets a–f); see Table 3[3]. These subsets can serve as a starting point for further research, for example, on writer-dependent recognition systems. However, except for subsets x, u, and a–f, no information is available correlating specific writers to filled forms.

[3] These subsets are marked in the image file name by an additional letter which corresponds to the set.

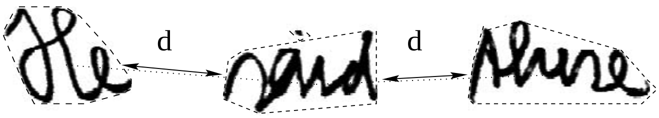


Fig. 7. Distance computation between connected components

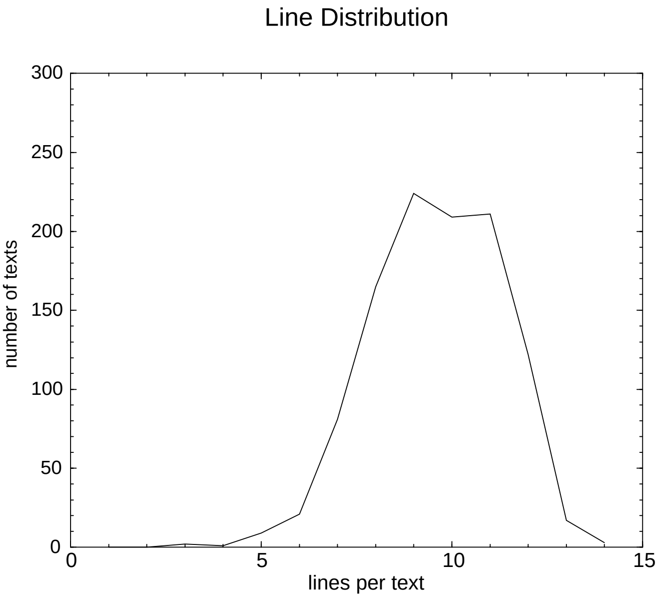


Fig. 8. Distribution of the number of handwritten text lines per text

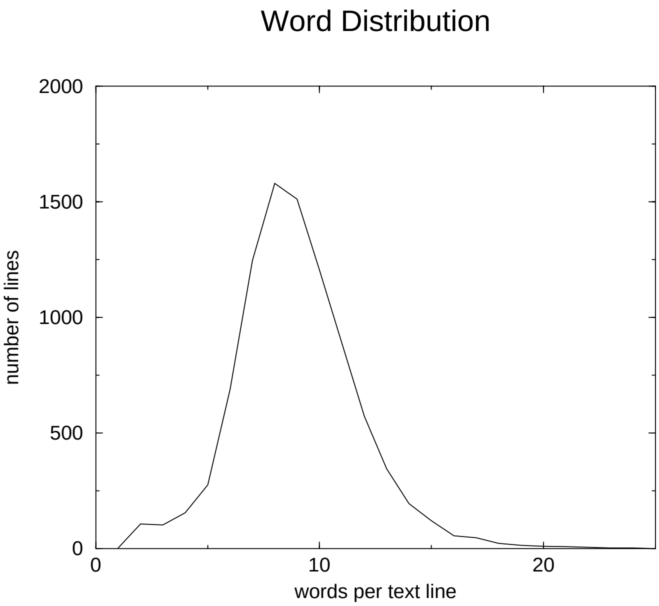


Fig. 9. Distribution of the number of handwritten words per text line

For the rest of the database there are between one and three forms from each writer. If there is more than one form, then these forms contain different text fragments, sometimes from different text categories. In most of the cases there is exactly one handwritten instance of a text fragment.

Most writers who contributed to the database are from Switzerland. However, we also received a number of forms from Greece and Germany. The database described in this paper was instrumental in our research on unconstrained English sentence recognition using statistical language models [12,16,17,13,14]. Research activities where the database is used by other groups are described in [8,1,24].

Table 2. Distribution of the forms over different texts and text classes

Text class	Number of forms related to the text number and class 1	2	3	4	5	6	Total
A	70	16	18	28	2	7	141
B	25	6	8	32	16	29	116
C	3	15	60	31	0	30	139
D	13	0	2	27	6	28	76
E	18	12	0	17	0	15	62
F	12	11	5	22	0	0	50
G	21	5	14	15	3	0	58
H	10	14	0	12	1	0	37
J	7	0	0	28	0	0	35
K	1	13	9	27	0	0	50
L	22	0	3	30	0	0	55
M	25	17	9	33	0	0	84
N	9	23	12	30	0	0	74
P	4	21	28	0	0	0	53
R	0	22	14	0	0	0	36
Total	240	175	182	332	28	109	1066

Table 3. Special subsets

Subset	Vocabulary	Number of forms	Number of lines	Text number	Number of writers
x	483	17	155	A01	10
u	765	29	287	A01	1
a	411	10	91	C03	1
b	411	10	89	C03	1
c	411	10	87	C03	1
d	411	10	97	C03	1
e	411	10	97	C03	1
f	411	9	94	C03	1

6 Conclusions

A database consisting of handwritten English sentences has been described in this paper. It is built upon the Lancaster-Oslo/Bergen corpus. The database can serve as a basis for research in handwriting recognition. In particular, it is potentially useful for recognition of general unconstrained English text utilizing knowledge beyond the lexicon level. Linguistic knowledge can be either supplied from external sources, or directly derived from the underlying corpus, which is available in electronic form.

A few preprocessing and segmentation procedures have been developed together with the database. They include skew normalization, extraction of the handwritten text from a form, and its segmentation into lines and individual words. These procedures were used as tools in the process of labeling the database. However, they can be integrated into any recognition system as well.

The version of the database described in this paper is freely available to other researchers upon request[4].

Acknowledgements. We would like to thank all the individuals who contributed to the database described in this paper. Particular thanks are due to Ergina Kavallieratou, Reto Rutsch, Alessandro Vinciarelli, Anita Barkensiek, and Stefan Knerr for helping collecting forms. Further thanks are due to the Swiss National Science Foundation for financial support (Project 2000-52'087.97).

References

1. A. Barkensiek, J. Rottland, A. Kosmala, G. Rigoll: Offline handwriting recognition using various hybrid modeling techniques and character n-grams. In: Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition, Amsterdam, The Netherlands, pp 343–352, (2000)
2. J.T. Favata, S.N. Srihari, V. Govindaraju: Off-line handwritten sentence recognition. In: A.C. Downton, S. Impedovo (eds) Progress in handwriting recognition. World Scientific, Singapore, pp 393–398, (1997)
3. W.N. Francis: Manual of information to accompany a standard sample of present-day edited American English for use with digital computers. Department of Linguistics, Brown University, Providence, Rhode Island, (1964)
4. I. Guyon, R.M. Haralick, J.J. Hull, I.T. Phillips: Database and benchmarking. In: H. Bunke, P.S.P. Wand (eds) Handbook of character recognition and document image analysis. World Scientific, Singapore, pp 779–799, (1997)
5. I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, S. Janet: Unipen project of on-line data exchange and benchmarks. In: Proc. 12th Int. Conf. on Pattern Recognition, Jerusalem, Israel, pp 29–33, (1994)
6. J.J. Hull: A database for handwritten text recognition research. IEEE Transn Pattern Anal Mach Intell 16(5):550–554, (1994)
7. S. Johansson, G.N. Leech, H. Goodluck: Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers. Department of English, University of Oslo, Oslo, (1978)
8. E. Kavallieratou, E. Stamatatos, N. Fakotakis, G. Kokkinakis: Handwritten character segmentation using transformation-based learning. In: Proc. 15th Int. Conf. on Pattern Recognition, Barcelona, Spain, 2:634–637, (2000)
9. D.H. Kim, Y.S. Hwang, S.T. Park, E.J. Kim, S.H. Paek, S.Y. Bang: Handwritten Korean character image database PE92. In: Proc. 2nd Int. Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan, pp 470–473, (1993)
10. G. Kim, V. Govindaraju, S.N. Srihari: Architecture for handwritten text recognition systems. In: S.-W. Lee (ed) Advances in handwriting recognition. World Scientific, Singapore, pp 163–172, (1999)
11. U. Mahadevan, R.C. Nagabushnam: Gap metrics for word separation in handwritten lines. In: Proc. 3rd Int. Conf. on Document Analysis and Recognition, Montréal, Canada, 1:124–127, (1995)
12. U. Marti: Offline Erkennung handgeschriebener Texte. PhD thesis, University of Bern, Switzerland, (2000)
13. U. Marti, H. Bunke: Handwritten sentence recognition. In: Proc. 15th Int. Conf. on Pattern Recognition, Barcelona, Spain, 3:467–470, (2000)
14. U. Marti, H. Bunke: Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system. To appear in Int J Pattern Recognition Artif Intell), (2001)
15. U.-V. Marti, H. Bunke: A full English sentence database for off-line handwriting recognition. In: Proc. 5th Int. Conf. on Document Analysis and Recognition, Bangalore, India, pp 705–708, (1999)
16. U.-V. Marti, H. Bunke: Towards general cursive script recognition. In: S.-W. Lee (ed) Advances in handwriting recognition. World Scientific, Singapore, pp 203–212, (1999)
17. U.-V. Marti, H. Bunke: Unconstrained handwriting recognition: language models, perplexity, and system performance. In: Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition, Amsterdam, The Netherlands, pp 463–468, (2000)
18. U.-V. Marti, R. Messerli, H. Bunke: Writer identification using text line-based features. (submitted)
19. N. Otsu: A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66, (1979)
20. T. Saito, H. Yamada, K. Yamamoto: On the data base ETL 9 of handprinted characters in JIS Chinese characters and its analysis. IEICE Trans J68-D(4):757–764, (1985)
21. G. Seni, E. Cohen: External word segmentation of off-line handwritten text lines. Pattern Recognition 27(1):41–52, (1994)
22. A.W. Senior, F. Fallside: Using unconstrained snakes for feature spotting in off-line cursive script. In: Proc. 2nd Int. Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan, pp 305–310, (1993)
23. C.Y. Suen, C. Nadal, R. Legault, T.A. Mai, L. Lam: Computer recognition of unconstrained handwritten numerals. Special Issue of Proc IEEE 7(80):1162–1180, (1992)
24. A. Vinciarelli, J. Luettin: Off-line cursive script recognition based on continous density hmm. In: Proc. 7th Int. Workshop on Frontiers in Handwriting Recognition, Amsterdam, The Netherlands, pp 493–498, (2000)
25. R.A. Wilkinson, J. Geist, S. Janet, P.J. Grother, C.J.C. Burges, R. Creecy, B. Hammond, J.J. Hull, N.W. Larsen, T.P. Vogl, C.L. Wilson: The first census optical character recognition systems conference. #NISTIR 4912, The U.S. Bureau of Census and the National Institute of Standards and Technology, Gaithersburg, Md., USA, (1992)

[4] Contact Horst Bunke at bunke@iam.unibe.ch