Lang Resources & Evaluation (2022) 56:1009–1022
https://doi.org/10.1007/s10579-022-09587-3

ORIGINAL PAPER - REPLICABILITY & REPRODUCIBILITY

A Spanish dataset for reproducible benchmarked offline handwriting recognition

Salvador España-Boquera[1] · Maria Jose Castro-Bleda[1]

Accepted: 17 February 2022 / Published online: 23 March 2022

Abstract In this paper, a public dataset for Offline Handwriting Recognition, along with an appropriate evaluation method to provide benchmark indicators at sentence level, is presented. This dataset, called SPA-Sentences, consists of offline handwritten Spanish sentences extracted from 1617 forms produced by the same number of writers. A total of 13,691 sentences comprising around 100,000 word instances out of a vocabulary of 3288 words occur in the collection. Careful attention has been paid to make the baseline experiments both reproducible and competitive. To this end, experiments are based on state-of-the-art recognition techniques combining convolutional blocks with one-dimensional Bidirectional Long Short Term Memory (LSTM) networks using Connectionist Temporal Classification (CTC) decoding. The scripts with the entire experimental setting have been made available. The SPA-Sentences dataset and its baseline evaluation are freely available for research purposes via the institutional University repository. We expect the research community to include this corpus, as is usually done with English IAM and French RIMES datasets, in their battery of experiments when reporting novel handwriting recognition techniques.

Keywords Handwriting recognition · Offline handwriting recognition · Datasets · Evaluation · Benchmarking · Experimental reproducibility · Spanish resources · Deep learning · Convolutional neural networks (CNN) · Long short term memory (LSTM) networks · Connectionist temporal classification (CTC)

✉ Maria Jose Castro-Bleda
mcastro@dsic.upv.es

Salvador España-Boquera
sespana@dsic.upv.es

[1] VRAIN Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Valencia, Spain

1 Introduction

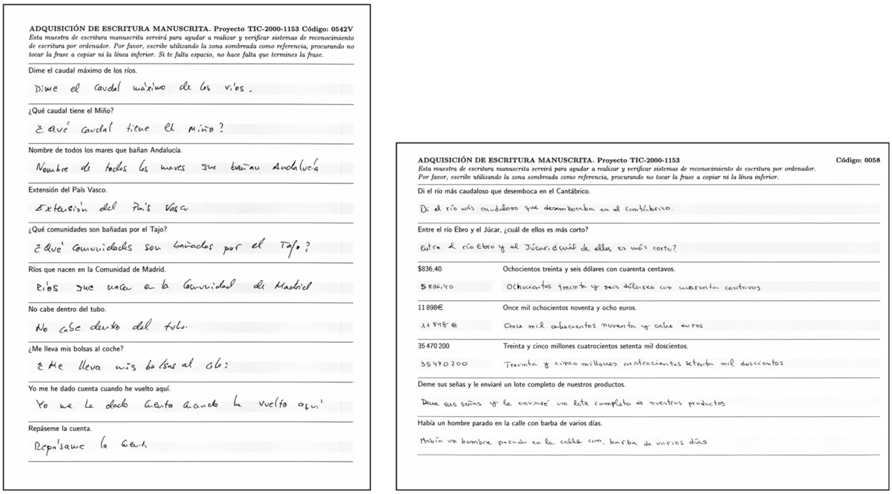
The availability of large amounts of data is a basic necessity for development, improvement, and assessment in all scientific research domains. Standard datasets make possible a fair comparison among different systems without bias. As in other scientific fields, having standard datasets has become an essential issue in the handwriting recognition research community. Most of these datasets have been developed for languages based on the letters of the classical Latin alphabet, although non-Latin script datasets would deserve a particular chapter (see Hussain et al., 2015 for a survey on this subject).

Focusing on the offline domain, the most widely used datasets include CEDAR (Hull, 1994), NIST (Wilkinson et al., 1992), MNIST (LeCun et al., 1998), CENPARMI (Suen et al., 1992), and IAM (Marti & Bunke, 2002). Not surprisingly, all of them are for modern English. It is more rare to find resources for other languages (the IRONOFF (Viard-Gaudin et al., 1999) and the RIMES (Grosicki et al., 2008) datasets are for modern French). Unfortunately, for Spanish, the third most used language in the world, there are very few resources. To the best of our knowledge, the only publicly available dataset for modern offline handwritten text in Spanish is described in (Juan et al., 2004), and it only comprises 485 images of numbers by 29 writers (2127 words), which is more than one order of magnitude smaller than the previously cited datasets. Other resources devoted to historical documents can be found in different ancient languages (see, for example, IAM-HistDB (Fischer et al., 2010), the Germana corpus (Pérez et al., 2009), or the tranScriptorium dataset (Sanchez et al., 2015); more in the survey (Hussain et al., 2015)).

This paper presents a novel comprehensive benchmark of a Spanish handwriting dataset aiming to alleviate difficulties in offline handwriting recognition and expand research in all aspects of Spanish script recognition. There were two main reasons to create this corpus. First of all, none of the above described Spanish datasets contains whole sentences. Secondly, although the set of Spanish graphemes is similar to the English set, some peculiarities may have to be considered (accented vowels, additional graphemes such as 'ñ', special symbols or abbreviations, ...).

The sentences of the dataset are chosen from different subtasks, such as numbers, questions, or general sentences. The entities at the lowest level are words that have been automatically segmented and manually checked for correctness.

It is a common practice, when releasing a corpus, to provide some standard partitions of training and test in order to make it easier for researchers working with it to report comparable figures of merit. It is also usual to provide a validation part from the training subset. Instead, we have provided five non-overlapping partitions of similar size as well as a proposal for performing *K*-fold cross validation experiments. That is, each experiment should be replicated five times, leaving aside a partition that should not be used at all except for a final evaluation stage. We have also proposed how to select a validation subset from the four remaining partitions devoted to training. When a *K*-fold scheme is not needed, the first partition can be the default split into training, validation, and test.



ADQUISICIÓN DE ESCRITURA MANUSCRITA. Proyecto TIC-2000-1153 Código: 0542V
Esta muestra de escritura manuscrita servirá para ayudar a realizar y verificar sistemas de reconocimiento de escritura por ordenador. Por favor, escribe utilizando la zona sombreada como referencia, procurando no tocar la frase a copiar ni la línea inferior. Si te falta espacio, no hace falta que termines la frase.

Dime el caudal máximo de los ríos.
Dime el caudal máximo de los ríos.

¿Qué caudal tiene el Miño?
¿Qué caudal tiene el Miño?

Nombre de todos los mares que bañan Andalucía.
Nombre de todos los mares que bañan Andalucía

Extensión del País Vasco.
Extensión del País Vasco

¿Qué comunidades son bañadas por el Tajo?
¿Qué comunidades son bañadas por el Tajo?

Ríos que nacen en la Comunidad de Madrid.
Ríos que nacen en la Comunidad de Madrid

No cabe dentro del tubo.
No cabe dentro del tubo.

¿Me lleva mis bolsas al coche?
¿Me lleva mis bolsas al coche?

Yo me he dado cuenta cuando he vuelto aquí.
Yo me he dado cuenta cuando he vuelto aquí

Repáseme la cuenta.
Repásame la cuenta

ADQUISICIÓN DE ESCRITURA MANUSCRITA. Proyecto TIC-2000-1153 Código: 0058
Esta muestra de escritura manuscrita servirá para ayudar a realizar y verificar sistemas de reconocimiento de escritura por ordenador. Por favor, escribe utilizando la zona sombreada como referencia, procurando no tocar la frase a copiar ni la línea inferior.

Di el río más caudaloso que desemboca en el Cantábrico.
Di el río más caudaloso que desembocaba en el Cantábrico.

Entre el río Ebro y el Júcar, ¿cuál de ellos es más corto?
Entre el río Ebro y el Júcar, ¿cuál de ellos es más corto?

$836,40 Ochocientos treinta y seis dólares con cuarenta centavos.
$836,40 Ochocientos treinta y seis dólares con cuarenta centavos

11 898€ Once mil ochocientos noventa y ocho euros.
11 898 € Once mil ochocientos noventa y ocho euros

35 470 200 Treinta y cinco millones cuatrocientos setenta mil doscientos.
35 470 200 Treinta y cinco millones cuatrocientos setenta mil doscientos

Deme sus señas y le enviaré un lote completo de nuestros productos.
Deme sus señas y le enviaré un lote completo de nuestros productos

Había un hombre parado en la calle con barba de varios días.
Había un hombre parado en la calle con. barba de varios días

Fig. 1 Two examples of filled acquisition forms for the SPA-Sentences dataset: the vertical form (left) contains 10 shorter sentences, while the number of lines in the horizontal form (right) is limited to 7, albeit a little wider

Finally, in order to provide baseline results for reference, some experiments with state-of-the-art techniques are reported. Special attention has been paid making these experiments both competitive and easily reproducible by choosing an out-of-the-box publicly available handwriting recognition engine, and providing the configuration parameters used in the proposed experiments.

The corpus and its baseline evaluation are freely available for research purposes (a nominal amount is charged for administration costs) via the institutional University repository at https://aplicat.upv.es/exploraupv/ficha-tecnologia/patente_software/27402?busqueda=spa-sentences. Commercial use requires a fee that varies depending on the type of business.

The rest of the paper is organized as follows. The next section describes the corpus in detail: its design, the acquisition and post-processing process, and some dataset statistics. Section 3 deals with the experimental setup, and Sect. 4 presents the recognition experiments. Some general conclusions are drawn in the final section.

2 The SPA-Sentences corpus

2.1 Corpus design and rationale

Our goal was to acquire a modern Spanish handwritten text dataset in the offline modality. As staff in a large University, we could ask many students to kindly and voluntarily collaborate with the corpus acquisition. This has allowed us to provide

Table 1 Number of sentences and words per subtask

Subtask	Sentences	Words	Vocabulary
Numbers	2313	18,698	104
Geographical queries	5790	46,112	247
Traveler questions	1362	11,085	645
General sentences	3012	25,522	2607
Total	12,477	101,417	3288

Note that the size of the vocabulary is lower than the sum of the different subtask vocabulary sizes due to common words

an extensive corpus with a large number of writers, which may hopefully capture a vast repertoire of writing styles.

We left the students to write with their pens not to impose any restrictions on the writing instrument. Hence, text written with different instruments is included in the dataset (mostly ink and ball-point pens). Another restriction was not to be too intrusive, annoying, or time-demanding for the volunteers. To this end, acquisition forms have been designed to fit in a one-sided A4 paper sheet. A short description of the purpose of the corpus and the acquisition procedure is included in the form header. An identification code is also included in the header to ease the post-processing of the filled and scanned forms.

Since we were mainly interested in handwritten recognition at the sentence level and not in document layout identification, paragraph detection, or text line extraction, we have decided to include horizontal rulers to simplify the line image extraction (see Fig. 1). In order to ease the writer's task, the form includes the typographic reference sentence and a guiding area to write into. Careful attention has been paid to limit the sentence to fit in a single line, which is not obvious since different people usually require different amounts of space to write the very same text. To cope with this issue, two different and complementary strategies have been applied:

– Two types of forms have been designed: portrait (vertical) and landscape (horizontal) forms (see also Fig. 1) in order to grasp a wider range of sentences: Longer sentences are collected into landscape forms to avoid getting compressed or deformed handwritten words while portrait forms, although only admit shorter sentences, may include 10 of them instead of 7. There is an average of 70 handwritten words per form in both cases.
– Nevertheless, volunteers were asked to stop writing if there was not enough space. This is not a serious issue because forms have been manually supervised afterward.

With no particular purpose at hand for this corpus (other than being general) and with the limitation of using short sentences, we have opted for combining four different subtasks to construct the written text:

Table 2 Distribution of portrait and landscape forms in each partition

Partition	Portrait		Landscape		Total	
	Forms	Lines	Forms	Lines	Forms	Lines
P0	160	1595	164	1145	324	2740
P1	160	1600	164	1144	324	2744
P2	160	1597	163	1138	323	2735
P3	159	1588	164	1145	323	2733
P4	160	1599	163	1140	323	2739
Total	1799	7979	818	5712	1617	13,691

The number of lines may be slightly lower than the corresponding number of lines per form (10 in portrait, 7 in landscape) multiplied by the number of forms since some lines have been removed in the post-processing step

- Numbers: different quantities of numbers and prices printed with digits and expressed as the quantity in letters. Prices are expressed in unities of euro and dollar, some with fractional parts.
- Geographical data queries extracted from Díaz-Verdejo et al. (1998).
- Traveler common questions extracted from Amengual et al. (2000).
- Unconstrained sentences were chosen to cover the possible lack of symbols, sequence of graphemes, and words not included in the other tasks.

An extensive set of different forms (1500) has been automatically generated. The number of sentences/lines and words of each subtask is summarized in Table 1.

2.2 Corpus acquisition and post-processing

The formatted documents were printed by an HP LaserJet 4100 DTN at a resolution of 600 dpi. The filled forms were scanned in gray level at 300 dpi. with a Hewlett Packard Scanjet ADF 6300c automatic sheet feeder scanner.

Printed forms were distributed among the staff in order to ask their students to voluntarily fill the forms at the beginning of a lecture. We initially believed that 1,500 forms were enough to not repeat them, but finally, 1617 forms (after removing problematic ones) were collected and scanned.

The initial version of this corpus (España Boquera et al., 2004) did not contain any specific partition on training, validation, and test subsets. In order to solve this issue, we have decided to divide it into five partitions. The number of forms in each partition is summarized in Table 2. In this way, *K*-fold cross validation experiments or classifier ensemble techniques can be easily designed using these partitions.

Scanned images have been cleaned and enhanced, while maintaining the gray level, by using a convolutional neural filter trained with some image pairs comprising clean scanned handwritten text (without the light gray boxes) and the same documents with these boxes overlapped.

Line extraction from filled forms can be easily performed using horizontal projection thanks to the rulers included in the forms. These rulers may be detected by computing the longest horizontal black run and horizontal projections. Skew and slant (Slavik & Govindaraju, 2001) were not corrected, mainly because skew does not seem an issue due to the use of rulers and the light gray area provided to the users. Relating the slant, we preferred to deliver the image lines *as is* so that researchers could try their preprocessing techniques.

Lines were segmented into words using a simple dynamic programming technique, considering that we had the corresponding text. We could have used an already trained recognition engine, but this basic approach turned out to work well in practice. The segmented lines were manually supervised to correct mistakes and, more importantly, to remove problematic filled forms.[1] Crossing outs have also been manually detected and annotated.

The final version of the corpus is delivered as a set of cleaned gray-scale images of filled forms in `png` format, together with a set of XML files describing their content. An example of a fragment of the XML file associated with the form of Fig. 1 (left) is illustrated in Fig. 2. As can be observed from the example, each page is divided into lines; each line is divided into the typographic part and the handwritten one; and, finally, each line is divided into words. Each part contains a label (in UTF-8 encoding) together with a bounding box. Typographic and handwritten parts have independent text labels allowing the format to cope with the case when the handwritten annotation has been modified to mark crossing outs and other issues.

The distribution of forms into five independent partitions is also provided. To this end, five index files indicate which files belong to each partition.

Finally, some Python scripts have been delivered to extract the text files and the image files associated with each line. Text lines are converted into a sequence of graphemes where special characters are replaced by labels, as illustrated in the following example (corresponding to the first line of the form illustrated in Fig. 1 (left) and whose XML file is shown in Fig. 2):

```
D i m e {space} e l {space} c a u d a l {space} m {a_acute} x i m o
 {space} d e {space} l o s {space} r {i_acute} o s {space} .
```

Images are extracted by cropping the cleaned page images using ImageMagick's convert tool.[2]

Finally, we have provided another Python script to resize all text line images to 96 pixels height while preserving the aspect ratio (using ImageMagick's convert). The chosen normalized height is roughly the median of the text line heights observed in the dataset.

[1] An example of a problematic filled form: writers had to copy questions such as "?‘En qué Comunidad desemboca el río Júcar?" (*In which community does the Júcar river flow into?*). Instead of doing so, some students wrote the response to the question (e.g. "En la Comunidad Valenciana" (*In the Valencian Community*)).

[2] https://imagemagick.org/script/convert.php.

```
<form name="v000_6"  code="0542V"
      image_name="v000_6.png" batch="000">
  <line line_number="0" task="GDQ">
    <typographic>
      <text>Dime el caudal máximo de los ríos .</text>
      <bndbox>
        <xmin>107</xmin>
        <ymin>496</ymin>
        <xmax>817</xmax>
        <ymax>541</ymax>
      </bndbox>
    </typographic>
    <handwritten>
      <text>Dime el caudal máximo de los ríos .</text>
      <bndbox>
        <xmin>142</xmin>
        <ymin>584</ymin>
        <xmax>2152</xmax>
        <ymax>677</ymax>
      </bndbox>
      <words>
        <word>
          <text>Dime</text>
          <bndbox>
            <xmin>144</xmin>
            <ymin>584</ymin>
            <xmax>2348</xmax>
            <ymax>677</ymax>
          </bndbox>
        </word>
        <word>
          <text>el</text>
          <bndbox>
           ...
            <xmin>926</xmin>
            <ymin>2897</ymin>
            <xmax>4450</xmax>
            <ymax>3039</ymax>
          </bndbox>
        </word>
      </words>
    </handwritten>
  </line>
</form>
```

Fig. 2 Example of the XML file corresponding to the scanned filled form of Fig. 1 (left)

2.3 Availability of SPA-Sentences dataset

The corpus is freely available for research purposes. Our University has an institutional repository to save the University community's production, personal or institutional, in collections. The researcher must access the associated link to the corpus in that repository https://aplicat.upv.es/exploraupv/ficha-tecnologia/patente_software/27402?busqueda=spa-sentences, and, after filling the form, the University will contact the researcher and release the SPA-Sentences dataset and its baseline evaluation.

3 Experimental setup

We have conducted a series of experiments to give some reference benchmarks for comparison purposes so that other researchers can have a baseline framework. We believe that it is not only essential to use state-of-the-art handwriting recognition techniques but also to make this experimentation as easily reproducible as possible.

3.1 State-of-the-art and reproducibility of experiments

Current state-of-the-art handwriting recognition techniques are mainly based on the Connectionist Temporal Classification (CTC) approach (Graves et al., 2006), which uses a particular RNN output layer and a loss function for sequence labeling tasks. CTC has been invariably used together with Long Short Term Memory (LSTM) networks (Gers et al., 2002; Hochreiter & Schmidhuber, 1997), either one dimensional (1D-LSTMs) (usually, Bidirectional LSTMs (BLSTMs) Graves and Schmidhuber (2005)) or multidimensional (MDLSTM) Graves and Schmidhuber (2009). CTC was first used for handwriting recognition in Graves et al. (2008). While previous works using CTC relied on handcrafted features (Doetsch et al., 2014; Graves et al., 2008), it is advantageous to combine the model with convolutional blocks to automatically learn the best features in an integrated way (Puigcerver, 2017; Shi et al., 2016). In order to cope with reproducibility, we have opted for using an out-of-the-box open-source handwriting recognition engine called PyLaia (Mocholí Calvo et al., 2018).

PyLaia is maintained as an open-source package under the MIT license and is available at https://github.com/jpuigcerver/PyLaia. It is based on Pytorch (Paszke et al., 2017) and it can be considered a successor of Laia (Puigcerver et al., 2016), which, likewise, was based on Torch (Collobert et al., 2011). This software has been extensively validated with experiments conducted on IAM (Marti & Bunke, 2002) and RIMES (Grosicki et al., 2008) datasets which are considered, as indicated in the introduction, the *de facto* standard for offline handwriting recognition evaluation on modern Latin script. The distribution of PyLaia provides some recipes for several corpora [e.g., IAM Marti and Bunke (2002), Cristo-Salvador Toselli et al. (2007), or Parzival Fischer et al. (2014)].

Although it is possible to use (Povey et al., 2011) to combine the output of the neural network with a language model, we have opted not to use any language model at all. Despite that, LSTMs are able to learn somewhat an implicit language model of grapheme sequences (Sabir et al., 2017).

3.2 Design of experiments

Experiments were conducted on computers equipped with a 6-core i5-8500 CPU at 3.00GHz and 8Gb of RAM running CentOS Linux release 7.5.1804. They were equipped with a GeForce GTX 1060 3GB GPU. The version of CUDA was V9.1.85, and cuDNN[3]Chetlur et al. (2014) was also used (version v5.1.10). The used PyLaia version was the `refactor_kws_egs_master` branch using the commit on Jun 5, 2019.

[3] https://developer.nvidia.com/cudnn.

All the experimentation setup parameters and training are identical to the recipe provided for the offline IAM dataset in the official PyLaia repository (Subfolder `egs/iam-htr` titled *Step-by-step Training Guide Using IAM Database*), even though that our corpus has a slightly larger number of graphemes (due to the presence of accented vowels and some letters and punctuation marks not present in the IAM corpus such as ‘ñ’, ‘Ñ’, ‘?‘’ or ‘!‘’). There are, nevertheless, some differences:

- Our corpus was preprocessed to make all text line images 96-pixel height, while the proposed IAM preprocessing scales lines to a height of 128 pixels. Consequently, the `fixed_input_height` parameter was reduced from 128 to 96.
- We have activated the option `use_baidu_ctc` (false by default) in order to use the Pytorch bindings for Baidu's Warp-CTC[4] (Amodei et al., 2016).
- There is an option for enabling the automatic generation of disturbed training patterns, as described in (Puigcerver, 2017). These distortions are computed on the fly and include rotations, translations, scaling, and shearing, as well as a gray-scale erosion and dilation. The original IAM recipe set this option to false. In our case, we have performed experiments with and without activating this option in order to measure the effect of this dynamic data augmentation technique, as was also done in Puigcerver (2017).
- The preprocessing stages proposed in the PyLaia IAM recipe were not applied, namely: enhancing the images by using the `imgtxtenh` tool[5], correcting the skew through ImageMagick's convert, removing all white borders from the images, and leaving a fixed size of 20 pixels on the left and the right sides of the image.

Our corpus is divided into five partitions (numbered from P0 to P4) to allow the use of *K*-fold cross validation.[6] Thus, the entire training and evaluation process was repeated five times. Firstly, P0 was used as the test set, the following part P1 was reserved for validation purposes, and the remaining parts were used for training. Regarding the four remaining partitions, the next partition used P1 as the test set, and so on.[7]

There is, in principle, a total of 5 × 2 different experiments (5 partitions and the presence/absence of distortions). However, we have decided to measure another feature: the distinction between accented and non-accented letters (which, in Spanish, is restricted to vowels, since the symbols ‘ñ’ and ‘Ñ’ are considered letters by themselves). There is a trade-off in this regard: on the one side, distinguishing accented vowels is the proper way of recognizing Spanish text, but, on the other side, some writers systematically skip the diacritical sign of the vowels. Tying both kinds of graphemes leads to a lower repertoire of labels, so it is a simpler task. To summarize, this leads to a total of $5 \times 2 \times 2 = 20$ different experiments.

The topology of the models used for all experiments is identical to that of the IAM recipe:

[4] https://github.com/baidu-research/warp-ctc.

[5] https://github.com/mauvilsa/imgtxtenh.

[6] Partition P0 is the default split if *K*-fold cross validation is not used.

[7] Circularly numbered, so that the next of P4 is P0.

Table 3 Number of epochs for training each experiment

ACC.	DIST.	Partition				
		P0	P1	P2	P3	P4
✓	✓	142	135	156	160	153
✓	×	133	129	197	106	201
×	✓	120	112	120	148	182
×	×	154	112	202	138	137

- There are 5 convolutional blocks of 16, 32, 48, 64 and 80 filters, respectively. All of them have kernels of size 3 × 3 pixels, with a stride of 1 pixel and a dilation of 1 pixel (that is, no dilation). All convolutional blocks are configured to use the Leaky Rectifier Linear Unit (LeakyReLU) activation function (Maas et al., 2013). The dropout probability was set to 0.[8] A MaxPooling of size 2 is applied only to the 3 first convolutional blocks. Batch normalization is not activated at any layer.
- Regarding the recurrent part of the model (bidirectional LSTMs), the number of hidden units in each direction is set to 256, and the number of recurrent blocks is set to 5, just as described in Puigcerver (2017) and Mocholí Calvo et al. (2018).

The learning rate was set to 0.0003, and the batch size was reduced to 8 due to memory restrictions of the GPU. Training has been configured to proceed until the model did not improve the results on validation for 20 epochs. These results are measured as the Character Error Rate (CER), although the Word Error Rate (WER) was also reported. Both measures were obtained with the `compute-wer` command provided by the Kaldi toolkit (Povey et al., 2011).

4 Experimental results

Each of the 20 different experiments was trained independently. The number of epochs in each configuration roughly varies between 100 and 200; values are detailed in Table 3. The training time required per epoch is around 390 seconds, while evaluating the validation set is near 37 seconds. This leads to a total training time between 12 and 23 hours, approximately, for each experiment. Figure 3 shows the evolution of the validation CER and WER for each training epoch for one of the experiments (partition P0 considering accents and distortions), although the same trend can be observed generally.

The validation CER and WER reached at the end of the training process are summarized in Tables 4 and 5, respectively. We can observe a tiny improvement when tying accented and non-accented vowels (option ACC. when accents are distinguished). A slight improvement is achieved when using the dynamic data augmentation to perturb the training image lines (option DIST. when distortions are applied). We believe that this trend should be extrapolated to the final results when

[8] We can observe that the similar recipe on Laia set the dropout of some convolutional layers to 0.2.

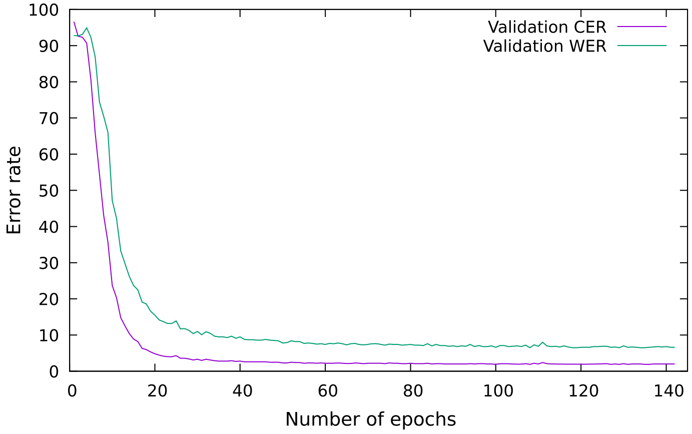


Fig. 3 Evolution of the validation CER and WER during training (in this case, for partition P0)

Table 4 Best CER on validation during training

ACC.	DIST.	Partition				
		P0	P1	P2	P3	P4
✓	✓	2.0	2.6	2.0	2.2	1.6
✓	×	2.5	2.8	2.2	2.7	1.6
×	✓	2.0	2.3	2.1	2.1	1.4
×	×	2.1	2.6	2.1	2.4	1.6

Table 5 Best WER on validation during training

ACC.	DIST.	Partition				
		P0	P1	P2	P3	P4
✓	✓	6.5	8.2	6.2	7.4	5.5
✓	×	8.2	8.8	7.1	8.7	5.7
×	✓	6.7	7.1	6.5	6.9	5.2
×	×	7.1	8.2	6.4	7.6	6.0

Table 6 CER on test

ACC.	DIST.	Partition					
		P0	P1	P2	P3	P4	Average
✓	✓	1.50	1.94	2.47	1.93	2.04	1.98
✓	×	1.85	2.37	2.66	2.36	2.27	2.30
×	✓	1.47	1.83	2.57	1.79	1.91	1.91
×	×	1.72	2.13	2.58	2.11	2.35	2.18

Table 7 WER on test

ACC.	DIST.	Partition					
		P0	P1	P2	P3	P4	Average
✓	✓	5.46	6.55	7.80	6.19	6.77	6.55
✓	×	6.59	7.85	8.33	7.79	7.57	7.63
×	✓	5.32	6.13	7.73	5.77	6.35	6.26
×	×	6.06	7.14	7.80	6.75	7.70	7.09

evaluating the test partitions. The CER and WER evaluated on these test partitions are shown in Tables 6 and 7. It is worth noticing that some results on the test are slightly better than on validation. In this regard, we can presume that some partitions may contain more difficult examples than others.

5 Conclusions

This paper presents a public dataset for offline modern handwriting recognition for the Spanish language. This dataset is quite extensive and was created by many writers. This fact contributes, in our opinion, to cope with a large variability of handwritten styles.

Although the corpus was acquired some time ago, only recently has it been delivered along with an easily reproducible and competitive state-of-the-art evaluation baseline, which may be very valuable for comparison purposes. The availability of the SPA-Sentences dataset, together with the baseline evaluation, should address the need of the research community interested in Spanish handwritten text recognition and should motivate the use of this corpus when measuring the quality of novel handwriting recognition techniques as is usually done now with the widely known IAM (Marti & Bunke, 2002) and the French RIMES (Grosicki et al., 2008) datasets.

Since the accompanying experimental results are easily reproducible, our future work includes a more extensive tuning of parameters to improve the reported figures of merit. Besides this parameter tuning, the use of preprocessing techniques such as those proposed by the out-of-the-box toolkit used in the reported experiments could be tried. Nevertheless, we believe that the text line images used *as is*, with no further preprocessing, have already reported very good results.

References

Amengual, J. C., Benedí, J. M., Casacuberta, F., Castaño, A., Castellanos, A., Jiménez, V. M., Llorens, D., Marzal, A., Prat, F., Vilar, J.M., Benedí, J.M., Casacuberta, F., Pastor, M., & Vidal. E. (2000). The EUTRANS-I speech translation system. *Machine Translation Journal, 15,* 75–103.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., Elsen, E., Engel, J., Fan, L., Fougner, C., Han, T., Hannun, A., Jun, B., LeGresley, P., Lin, L., Narang, S., Ng, A., Ozair, S., Prenger, R., Raiman, J., Satheesh, S., Seetapun, D., Sengupta, S., Wang, Y., Wang, Z., Wang, C., Xiao, B., Yogatama, D., Zhan, J., & Zhu. Z.

(2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. In *Proceedings of the 33rd international conference on international conference on machine learning (ICML)* (Vol. 48, pp. 173–182). JMLR.org.

Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., & Shelhamer, E. (2014). cuDNN: Efficient primitives for deep learning. CoRR **abs/1410.0759**. http://arxiv.org/abs/1410.0759.

Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). Torch7: A Matlab-like environment for machine learning. In *Proceedings of big learning 2011: NIPS 2011 workshop on algorithms, systems, and tools for learning at scale*.

Díaz-Verdejo, J. E., Peinado, A. M., Rubio, A. J., Segarra, E., Prieto, N., & Casacuberta, F. (1998). ALBAYZIN: A task-oriented Spanish speech corpus. In *Proceedings of the first international conference on language resources and evaluation (LREC)* (pp. 497–501). Granada, Spain.

Doetsch, P., Kozielski, M., & Ney, H. (2014). Fast and robust training of recurrent neural networks for offline handwriting recognition. In *Proceedings of the 14th international conference on frontiers in handwriting recognition (ICFHR)* (pp. 279–284). IEEE.

España Boquera, S., Castro Bleda, M. J., & Hidalgo, J. L. (2004). The SPARTACUS-Database: A Spanish sentence database for offline handwriting recognition. In *Proceedings of the fourth international conference on language resources and evaluation (LREC)* (pp. 227–230). Lisbon, Portugal.

Fischer, A., Baechler, M., Garz, A., Liwicki, M., & Ingold, R. (2014). A combined system for text line extraction and handwriting recognition in historical documents. In *Proceedings of the 11th IAPR international workshop on document analysis systems (DAS)* (pp. 71–75). IEEE.

Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., & Stolz, M. (2010). Ground Truth Creation for Handwriting Recognition in Historical Documents. In *Proceedings of the 9th IAPR international workshop on document analysis systems (DAS)* (pp. 3–10). ACM, New York, NY, USA. https://doi.org/10.1145/1815330.1815331.

Gers, F. A., Schraudolph, N. N., & Schmidhuber, J. (2002). Learning precise timing with LSTM recurrent networks. *Journal of machine learning research, 3*(Aug), 115–143.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on machine learning (ICML)* (pp. 369–376). ACM.

Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., & Schmidhuber, J. (2008). A novel connectionist system for unconstrained handwriting recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 31*(5), 855–868.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18*(5–6), 602–610.

Graves, A., & Schmidhuber, J. (2009). Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pp. 545–552.

Grosicki, E., Carré, M., Brodin, J. M., & Geoffrois, E. (2008). RIMES evaluation campaign for handwritten mail processing. In *Proceedings of the 11th international conference on frontiers in handwriting recognition (ICFHR)*, pp. 1–6. Concordia University, Montreal, Canada. https://hal.archives-ouvertes.fr/hal-01395332.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 16*(5), 550–554.

Hussain, R., Raza, A., Siddiqi, I., Khurshid, K., & Djeddi, C. (2015). *A comprehensive survey of handwritten document benchmarks: structure, usage and evaluation* (p. 46). Image and Video Processing: EURASIP J.

Juan, A., Toselli, A. H., Domnech, J., González, J., Salvador, I., Vidal, E., & Casacuberta, F. (2004). Integrated handwriting recognition and interpretation via finite-state models. *International Journal of Pattern Recognition and Artificial Intelligence, 18*(04), 519–539.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE, 86*(11), 2278–2324

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of the international conference on international conference on machine learning (ICML)* (Vol. 30, p. 3).

Marti, U. V., & Bunke, H. (2002). The IAM-database: An English sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition, 5*, 39–46.

Mocholí Calvo, C., Mocholí-CalvoMocholí-Calvo, C.Tutored by E. VIdal and J. Puigcerver. (2017–2018). Development and experimentation of a deep learning system for convolutional and recurrent neural networks. Master's thesis, ETSINF Universitat Politècnica de València, Valencia (Spain).

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., & Lerer, A. (2017). Automatic differentiation in PyTorch. In *Proceedings of the 31st conference on neural information processing systems (NIPS)*. Long Beach, CA, USA.

Pérez, D., Tarazón, L., Serrano, N., Castro, F., Terrades, O.R., & Juan-Císcar, A. (2009). The GERMANA database. In *10th International conference on document analysis and recognition* (pp. 301–305).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). *The Kaldi speech recognition toolkit*. Technical report: IEEE signal processing society.

Puigcerver, J. (2017). Are multidimensional recurrent layers really necessary for handwritten text recognition? In *Proceedings of the 14th IAPR international conference on document analysis and recognition (ICDAR)* (Vol. 01, pp. 67–72). https://doi.org/10.1109/ICDAR.2017.20.

Puigcerver, J., Martin-Albo, D., & Villegas, M. (2016). Laia: A deep learning toolkit for HTR.

Sabir, E., Rawls, S., & Natarajan, P. (2017). Implicit language model in LSTM for OCR. In *Proceedings of the 14th IAPR international conference on document analysis and recognition (ICDAR)* (Vol. 7, pp. 27–31). IEEE.

Sanchez, J. A., Toselli, A. H., Romero, V., & Vidal, E. (2015). ICDAR 2015 competition HTRtS: Handwritten text recognition on the tranScriptorium dataset. In *Proceedings of the 13th international conference on document analysis and recognition (ICDAR)*.

Shi, B., Bai, X., & Yao, C. (2016). An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 39*(11), 2298–2304.

Slavik, P., & Govindaraju, V. (2001). Equivalence of Different Methods for Slant and Skew Corrections in Word Recognition Applications. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 23*(3), 323–326.

Suen, C. Y., Nadal, C., Legault, R., Mai, T. A., & Lam, L. (1992). Computer recognition of unconstrained handwritten numerals. *Special Issue of Proceedings of IEEE, 7*(80), 1162–1180.

Toselli, A. H., Romero, V., & Vidal, E. (2007). Viterbi based alignment between text images and their transcripts. In *Proceedings of the workshop on language technology for cultural heritage data (LaTeCH)* (pp. 9–16).

Viard-Gaudin, C., Lallican, P. M., Knerr, S., & Binter, P. (1999). The IRESTE on/off (IRONOFF) dual handwriting database. In *Proceedings of the fifth international conference on document analysis and recognition (ICDAR) (pp. 455–458)*. Bangalore, India.

Wilkinson, R., Geist, J., Janet, S., Grother, P., Burges, C., Creecy, R., Hammond, B., Hull, J., Larsen, N., Vogl, T., & Wilson, C. (1992). The first census optical character recognition systems conference. In *#NISTIR 4912. The U.S. Bureau of Census and the National Institute of Standards and Technology*, Gaithersburg, MD.