Lang Resources & Evaluation manuscript No.
(will be inserted by the editor)

A Spanish dataset for reproducible benchmarked offline handwriting recognition

Salvador España-Boquera, Maria Jose Castro-Bleda

Received: date / Accepted: date

Abstract In this paper, a public dataset for Offline Handwriting Recognition, along with an appropriate method of evaluation to provide benchmark indicators at sentence level, is presented. This dataset, called SPA-Sentences, consists of offline handwritten Spanish sentences extracted from 1,617 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 3,288 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 and the scripts with the entire experimental setting have been made available. This release, together with the baseline evaluation, should motivate 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)

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 therefore become an important issue

S. España-Boquera
VRAIN Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Valencia, Spain
E-mail: sespana@dsic.upv.es

M.J. Castro-Bleda
VRAIN Valencian Research Institute for Artificial Intelligence, Universitat Politècnica de València, Valencia, Spain
E-mail: mcastro@dsic.upv.es (Corresponding Author)

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 special chapter (see [18] for a survey on this subject).

Focusing on the offline domain, most widely used datasets include CEDAR [17], NIST [36], MNIST [20], CENPARMI [33] and IAM [22]. Not surprisingly, all of them are for modern English. It is more rare to find resources for other languages (the IRONOFF [35] and the RIMES [15] 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 [19] and it only comprises 485 images of numbers by 29 writers (2,127 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 [9], the Germana corpus [25] or the tranScriptorium dataset [30]; more in survey [18]).

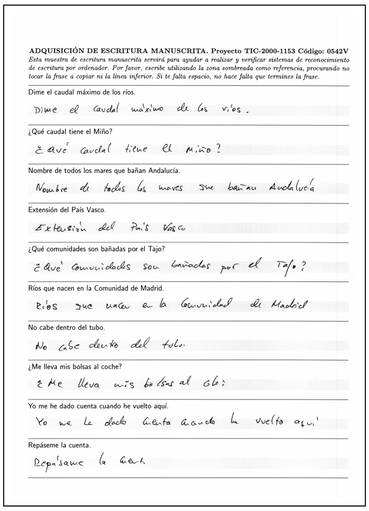
This paper presents a novel comprehensive benchmark of a Spanish handwriting dataset aiming to alleviate difficulties in offline handwriting recognition and to expand research in all aspects of Spanish script recognition. The SPA-Sentences dataset is made of sentences and there were two main reasons to create this corpus. First of all, none of the above described Spanish datasets contains full sentences. Secondly, although the set of Spanish graphemes is similar to the English set, there are some pecularities that may have to be taken into consideration (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 which 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 which should not be used at all excepting for a final evaluation stage. We have also proposed how to select a validation subset from the four remaining partitions devoted to training.

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

The rest of the paper is organized as follows. Next section describes the corpus in detail: its design, the acquisition and post-processing process, along with some statistics of the corpus. Section 3 deals with the experimental setup and Section 4 presents the recognition experiments. Some general conclusions are drawn in a final section. The corpus is freely available upon request for research purposes.



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.

¿Qué caudal tiene el Miño?

Nombre de todos los mares que bañan Andalucía.

Extensión del País Vasco.

¿Qué comunidades son bañadas por el Tajo?

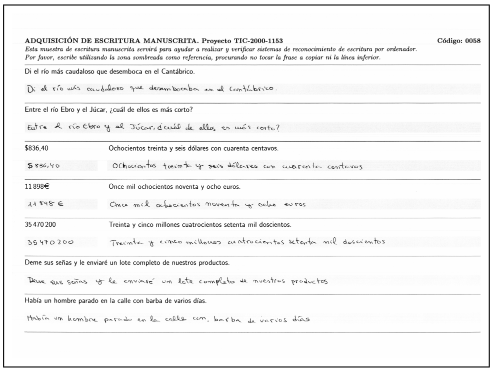
Ríos que nacen en la Comunidad de Madrid.

No cabe dentro del tubo.

¿Me lleva mis bolsas al coche?

Yo me he dado cuenta cuando he vuelto aquí.

Repáseme 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.

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

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

11 898€ Once mil ochocientos noventa y ocho euros.

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

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

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.

2 The SPA-Sentences corpus

2.1 Corpus design and rationale

Our goal was to acquire a dataset of modern Spanish handwritten text in the offline modality. As staff in a large University, we could ask a large group of students to kindly and voluntarily collaborate with this corpus acquisition. This has allowed us to provide an extensive corpus with a large number of writers, which may hopefully capture a wide repertoire of writing styles. Nevertheless, this variability is biased by the fact that all subjects came from a similar background and also by the fact that this text is copied and not spontaneously written or written in a different range of situations.

We left the students to write with their own pens in order not to impose any restriction 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 one sided A4 paper sheet. A short description of the purpose of the corpus and the acquisition procedure is included in the header of the form. An identification code is also included in the header in order 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 Figure 1). In order to ease the writer task, the typographic reference sentence and a guiding area to write into appear between two rulers. Careful attention has been paid to limit the sentence length to make them fit in a single line, which is not at all obvious since different people usually require a different

Table 1 Number of sentences and words per subtask.

Subtask	Sentences	Words	Vocabulary
Numbers	2,313	18,698	104
Geographical queries	5,790	46,112	247
Traveler questions	1,362	11,085	645
General sentences	3,012	25,522	2,607
Total[a]	12,477	101,417	3,288

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

amount 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 Figure 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 where asked to stop writing if there was not enough space. This is not a serious issue because forms have been manually supervised afterwards.

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:

– 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 of them with fractional parts.
– Geographical data queries extracted from [5].
– Traveler common questions extracted from [1].
– Unconstrained sentences chosen to cover the possible lack of symbols, sequence of graphemes, and words not included in the other tasks.

A large set of different forms (1,500) has been automatically generated. The number of sentences/lines and words of each subtask, in the final set of forms, is summarized in Table 1.

2.2 Corpus acquisition and post-processing

The formatted documents were printed by a 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 scanner with automatic sheet feeder.

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 in order not to repeat them, but finally 1,617 (after removing problematic ones) have been collected and scanned.

Table 2 Distribution of portrait and landscape forms in each partition.

	Portrait		Landscape		Total	
Partition	forms	lines[a]	forms	lines	forms	lines
P0	160	1,595	164	1,145	324	2,740
P1	160	1,600	164	1,144	324	2,744
P2	160	1,597	163	1,138	323	2,735
P3	159	1,588	164	1,145	323	2,733
P4	160	1,599	163	1,140	323	2,739
Total	1,799	7,979	818	5,712	1,617	13,691

[a] 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.

The initial release of this corpus [7] 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 5 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 by 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. The skew and the slant [32] have not been corrected, mainly because the 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 research groups can try their own preprocessing techniques.

Lines were segmented into words by means of a simple dynamic programming technique, taking into account 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 have been manually supervised in order 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 grasycale 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 to the form of Figure 1 (left) is illustrated in Figure 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

[1] As an example of problematic filled form, writers have to copy questions such as "¿En qué Comunidad desemboca el río Júcar?" (*In which community does the Júcar river flow into?*) but, some of them, filled the form with the response to the question instead of the question itself (e.g. "En la Comunidad Valenciana" (*In the Valencian Community*)).

with the case when the handwritten annotation has been modified to mark crossing outs and other issues.

A 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 to 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 Figure 1 (left) and whose XML file is shown in Figure 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 by using ImageMagick's convert tool.[2]

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

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 important 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 use of the Connectionist Temporal Classification (CTC) approach [11], which makes use of 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 [16,10] either one dimensional (1D-LSTMs) (usually, Bidirectional LSTMs (BLSTMs) [13]) or multidimensional (MDLSTM) [14]. CTC was first used for handwriting recognition in [12]. While previous works using CTC relied on handcrafted features [12,6], it is very advantageous to combine the model with convolutional blocks in order to automatically learn the best features in an integrated way [31,27]. In order to cope with reproducibility, we have opted for using an out-of-the-box ready to use open source hanwriting recognition engine called PyLaia [23].

PyLaia is mantained as an open-source package under the MIT license and is available at `https://github.com/jpuigcerver/PyLaia`. It is based on Pytorch [24] and can be considered as a successor of Laia [28] which, likewise, was based on

[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 Figure 1 (left).

Torch [4]. This software has been extensively validated with experiments conducted on IAM [22] and RIMES [15] databases which are considered, as indicated in the introduction, the *de facto* standard for offlline handwriting recognition evaluation on modern Latin script. The distribution of PyLaia provides some recipes for several corpus (e.g. IAM [22], Cristo-Salvador [34] or Parzival [8]).

Although it is possible to use [26] in order to combine the output of the neural network with a language model, we have opted for not using any language model at all. In spite of that, LSTMs are able to learn somewhat an implicit language model of grapheme sequences [29].

3.2 Design of experiments

Experiments have been 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 are equipped with a GeForce GTX 1060 3GB GPU. The version of CUDA was V9.1.85 and cuDNN[3] [3] 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.

All the experimentation setup parameters and training are identical to the recipe provided for the offline IAM database in the official PyLaia repository (Subfolder `egs/iam-htr` titled *Step-by-step Training Guide Using IAM Database*) in spite of the fact 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 has been 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 has been 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] [2].
- There is an option for enabling the automatic generation of disturbed training patterns, as described in [27]. 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 [27].
- The preprocessing stages proposed in the PyLaia IAM recipe have not beed applied, namely: enhancing the images by using the `imgtxtenh` tool[5], correcting the skew by means of 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 5 partitions (numbered from P0 to P4) to allow the use of K-fold cross validation. Thus, the entire training and evaluation process has been repeated 5 times: a partition has been left aside during training and has only been used at the end to report the final test result. Regarding the four remaining partitions, the next partition[6] has been reserved for validation purposes, while three remaining ones have been used for training.

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, namely: 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

[3] `https://developer.nvidia.com/cudnn`
[4] `https://github.com/baidu-research/warp-ctc`
[5] https://github.com/mauvilsa/imgtxtenh
[6] In a circular way, so that the next of P4 is P0.

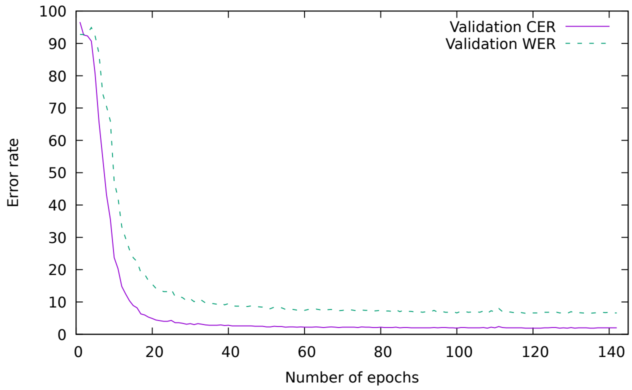


Fig. 3 Evolution of the validation CER and WER during training (in this case, for partition P0).

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:

- There are 5 convolutional blocks 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 Units (LeakyReLU) activation function [21]. The dropout probability has been set to 0.[7] 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 [27] and in [23].

The learning rate has been set to 0.0003 and the batch size has been reduced to 8 due to the memory restrictions of the GPU. Training has been configured to proceeded 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) is also reported. Both measures have been obtained with the `compute-wer` command provided by the Kaldi toolkit [26].

4 Experimental results

Each of the 20 different experiments has been 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 the time for 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 in general.

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

Table 3 Number of epochs for training each experiment.

		Partition				
ACC.	DIST.	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

Table 4 Best CER on validation during training.

		Partition				
ACC.	DIST.	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.

		Partition				
ACC.	DIST.	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.

		Partition					
ACC.	DIST.	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.

		Partition					
ACC.	DIST.	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

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 little 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 consisting on perturbing the training image lines (option DIST. when distortions are applied). We believe that this trend should be extrapolable to the final results when 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 test are slightly better than on validation. In this regard, we can presume that some partitions may simply contain more difficult examples than others.

5 Conclusions

In this paper a public dataset for offline modern handwriting recognition for the Spanish language is presented. This dataset is quite extensive and has been produced by a large number of 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 [22] and the French RIMES [15] datasets.

Since the accompanying experimental results are easily reproducible, our future work includes a larger tuning of parameters in order 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

1. Amengual, J.C., Benedí, J.M., Casacuberta, F., Castaño, A., Castellanos, A., Jiménez, V.M., Llorens, D., Marzal, A., Pastor, M., Prat, F., Vidal, E., Vilar, J.M.: The EUTRANS-I speech translation system. Machine Translation Journal **15**, 75–103 (2000)
2. Amodei, D., et al.: 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 (2016)
3. Chetlur, S., Woolley, C., Vandermersch, P., Cohen, J., Tran, J., Catanzaro, B., Shelhamer, E.: cuDNN: Efficient primitives for deep learning. CoRR **abs/1410.0759** (2014). URL `http://arxiv.org/abs/1410.0759`
4. Collobert, R., Kavukcuoglu, K., Farabet, C.: 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 (2011)
5. Díaz-Verdejo, J.E., Peinado, A.M., Rubio, A.J., Segarra, E., Prieto, N., Casacuberta, F.: 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 (1998)
6. Doetsch, P., Kozielski, M., Ney, H.: 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 (2014)
7. España Boquera, S., Castro Bleda, M.J., Hidalgo, J.L.: 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 (2004)
8. Fischer, A., Baechler, M., Garz, A., Liwicki, M., Ingold, R.: 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 (2014)
9. Fischer, A., Indermühle, E., Bunke, H., Viehhauser, G., Stolz, M.: 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 (2010). DOI 10.1145/1815330.1815331. URL `http://doi.acm.org/10.1145/1815330.1815331`

10. Gers, F.A., Schraudolph, N.N., Schmidhuber, J.: Learning precise timing with LSTM recurrent networks. Journal of machine learning research **3**(Aug), 115–143 (2002)
11. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: 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 (2006)
12. Graves, A., Liwicki, M., Fernández, S., Bertolami, R., Bunke, H., Schmidhuber, J.: A novel connectionist system for unconstrained handwriting recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **31**(5), 855–868 (2008)
13. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks **18**(5-6), 602–610 (2005)
14. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: Advances in Neural Information Processing Systems, pp. 545–552 (2009)
15. Grosicki, E., Carré, M., Brodin, J.M., Geoffrois, E.: 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 (2008). URL `https://hal.archives-ouvertes.fr/hal-01395332`
16. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
17. Hull, J.J.: A database for handwritten text recognition research. IEEE Trans. on Pattern Analysis and Machine Intelligence **16**(5), 550–554 (1994)
18. Hussain, R., Raza, A., Siddiqi, I., Khurshid, K., Djeddi, C.: A comprehensive survey of handwritten document benchmarks: structure, usage and evaluation. EURASIP J. Image and Video Processing **46** (2015)
19. Juan, A., Toselli, A.H., Domnech, J., González, J., Salvador, I., Vidal, E., Casacuberta, F.: Integrated Handwriting Recognition and Interpretation via Finite-State Models. Int. J. Patt. Recognition and Artifial Intelligence **18**(04), 519–539 (2004)
20. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proc. IEEE 86 (11), pp. 2278—2324 (1998)
21. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: Proceedings of the International Conference on International Conference on Machine Learning (ICML), vol. 30, p. 3 (2013)
22. Marti, U.V., Bunke, H.: The IAM-database: an English sentence adatabase for offline handwriting recognition. International Journal on Document Analysis and Recognition **5**, 39–46 (2002)
23. Mocholí Calvo, C., et al.: Development and experimentation of a deep learning system for convolutional and recurrent neural networks. Master's thesis, ETSINF Universitat Politècnica de València, Valencia (Spain) (2017-2018)
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS). Long Beach, CA, USA (2017)
25. Pérez, D., Tarazón, L., Serrano, N., Castro, F., Terrades, O.R., Juan-Císcar, A.: The GERMANA Database. 10th International Conference on Document Analysis and Recognition pp. 301–305 (2009)
26. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The Kaldi speech recognition toolkit. Tech. rep., IEEE Signal Processing Society (2011)
27. Puigcerver, J.: 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 (2017). DOI 10.1109/ICDAR.2017.20
28. Puigcerver, J., Martin-Albo, D., Villegas, M.: Laia: A deep learning toolkit for HTR (2016)
29. Sabir, E., Rawls, S., Natarajan, P.: 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 (2017)
30. Sanchez, J.A., Toselli, A.H., Romero, V., Vidal, E.: ICDAR 2015 competition HTRtS: Handwritten Text Recognition on the tranScriptorium dataset. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR) (2015)

31. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence **39**(11), 2298–2304 (2016)
32. Slavik, P., Govindaraju, V.: Equivalence of Different Methods for Slant and Skew Corrections in Word Recognition Applications. IEEE Trans. on Pattern Analysis and Machine Intelligence **23**(3), 323–326 (2001)
33. Suen, C.Y., Nadal, C., Legault, R., Mai, T.A., Lam, L.: Computer recognition of unconstrained handwritten numerals. Special Issue of Proc. IEEE **7**(80), 1162–1180 (1992)
34. Toselli, A.H., Romero, V., Vidal, E.: 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 (2007)
35. Viard-Gaudin, C., Lallican, P.M., Knerr, S., Binter, P.: 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 (1999)
36. Wilkinson, R., Geist, J., Janet, S., Grother, P., Burges, C., Creecy, R., Hammond, B., Hull, J., Larsen, N., Vogl, T., Wilson, C.: 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 (1992)