Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 1, March, 2026, pp. 31-38
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 107

Toolkit for Generating and Augmenting Hungarian Handwritten Text Recognition Dataset

Mohammed A.S Al-Hitawi[1], Ali Layth Alzaidy[2], M A Alazaizi [3], Mustafa Adnan Tharthar [4,]
[1]Department of artificial intelligence, College of Information Technology, University of Fallujah, 31002, Fallujah, Iraq
[2]Department of religious education and Islamic studies, Iraqi Sunni Affairs, Baghdad, Iraq
[3] Faculty of Management, Technische Universität München (TUM), Arcisstraße 21, 80333 Munich, Germany
[4] Registration and Students Affairs, University Headquarter, University of Anbar, 31001, Ramadi, Anbar, Iraq

Article Info

Article history:

Received Sept., 27, 2025
Revised Dec. 20, 2025
Accepted Feb., 29, 2026

Keywords:

Handwritten text recognition
Hungarian HTR dataset
Deep learning
Culture heritages
Digitization
HTR toolkit
Machine learning

ABSTRACT

Handwritten Text Recognition (HTR) in low-resource languages, such as Hungarian, faces persistent challenges due to the limited availability of high-quality datasets. This paper introduces the HuHTR-Toolkit, an open-source framework for generating and augmenting Hungarian HTR datasets. The toolkit generates realistic handwritten text images from open-source corpora, enabling the creation of large-scale datasets and offering a wide range of augmentation techniques including mathematical, chromatic, and style-based transformations to enhance model robustness. Using the Belval HuHTR-Toolkit, we produced over three million synthetic text-image line pairs and evaluated their impact on transformer-based models. Data augmentation, the second approach to expanding data, applies image-processing techniques. Experimental results indicate that synthetic data with touch augmentation greatly improves character and word-level accuracy, reducing reliance on costly human-annotated datasets. The toolkit and generated datasets are publicly available to support further research in low-resource HTR and cultural heritage digitization.

Corresponding Author:

Mohammed A.S. Al-Hitawi
Department of Artificial Intelligence, College of Information Technology, University of Fallujah
31002, Fallujah, Iraq
Email: al_hitawe@uofallujah.edu.iq

1. INTRODUCTION

The technology behind Optical Character Recognition (OCR) is advanced, but Handwritten Text Recognition (HTR) remains in its development stages [1]. OCR has currently accomplished very effectively for the English language, but there are some obstacles and high error rates for other languages besides English including Arabic because of mixed letters in expressions (there is no a forum for most of the letters). Additionally the Hungarian has a few special characters, so our plan on tackling the gap for offline handwriting recognition for the Hungarian language acquisition, considered perhaps the most prevalent machine learning difficulties by introducing collection of synthetics and augmentation dataset. Hungarian uses a distinct script, along with its unique writing style. Neural network models require large amounts of data, particularly when using attention mechanisms. To address this, new data were generated for training, which can be applied in two stages: first, pre-training on artificially generated data, and second, fine-tuning on human annotated samples. Accordingly, we produced a new synthetic Hungarian dataset by adapting an existing technology originally developed for another language, HuTRDG[1].

Aim and Contributions

This study addresses the scarcity of Hungarian HTR datasets by introducing HuHTR-Gen, an open-source toolkit for generating and augmenting synthetic handwriting data. The toolkit provides a user-friendly platform for dataset creation, supports diverse handwriting styles and augmentations, and enables research on Hungarian handwriting recognition.

1. A scalable toolkit for creating Hungarian handwritten datasets with support for diacritical marks and 200+ open-source fonts.
2. Integration of augmentation techniques (morphological, chromatic, geometric) to increase dataset diversity.
3. Release of a publicly available dataset containing over 3M synthetic line-level pairs for Hungarian HTR.
4. Empirical evaluation showing improvements in CER and WER when training transformer-based models with synthetic + real data.

2. LITERATURE REVIEW

2.1. Related work

Previous research efforts have explored several approaches to handwritten text recognition. One particularly challenging task in synthesizing digital information involved the use of nearest-neighbour-based collection OCR [10]. However, there is a limited body of literature specifically addressing Hungarian Handwritten Text Recognition (HTR) and the generation or augmentation of synthetic datasets. Historical Hungarian scripts include Textualis (late 11th to 13th centuries) and Cursive (13th to 15th centuries). Most models were designed using a Convolutional Recurrent Neural Network (CRNN) combined with Connectionist Temporal Classification (CTC) loss. The hypotheses were taught and evaluated using three existing ground-truth corpora: the e-NDP corpus, the Alcar-HOME database, and the Himanis project, totalling 120k lines of text and nearly a million tokens. This study explains the training architecture and corpora for a worker, as well as the fundamental training barriers, outcomes, and prospective utilization of HTR approaches on medieval chronicle texts [2]. Highlighting an ancient Hungarian handwritten sample with observations. Although OCR of superior digital files is an established area that utilizes numerous commercial apps and extensive files of the material in the wild, no available data sets can be used to create and validate keep track of OCR strategies that are powerful to non-uniform illumination image blur, strong noise, built-in de-noising, sharpening, contraction, and other remnants observed in many photographs taken with mobile phones and tablets. This dataset contains 2,113 distinct pages from scientific articles, captured by multiple users with 23 different cell phones. It includes accurate coordinates and textual annotations for approximately 500,000 text lines, complementing the 19,728 images of varying visual quality [3]. Although not Hungarian-specific, the study describes synthetic data extraction for HTR, which can be adapted for Hungarian handwriting recognition. Research has demonstrated the effectiveness of encoder–decoder-based approaches in identifying and correcting OCR errors in documents, particularly when combined with Context-based Character Correction (CCC) models. Ongoing work also explores experimentation on the Gold benchmark corpus. In the case of Hungarian, OCR performance can be strengthened through the use of artificially generated training data, with recent studies investigating synthetic data augmentation for Hungarian OCR [4]. Generative Adversarial Networks (GAN) is an approach for creating unique generated examples that does not need any prior understanding of the likely differences between illustrations. In the event of a test set with only a few thousand labelled instances described in a space of high dimensions, the classifier may perform badly. Configure a GAN to create generated images to augment image or signal databases to improve the effectiveness of classifiers. The technique has been tested on a variety of samples containing handwritten digits, including Latin, which the Hungarian speech is an outgrowth of [5]. Artificial data may have been used for generating or enhancing the currently available training dataset, improving recognition performance. Any public digital data from online newspapers and other locations can be utilized for producing synthetic data while preserving the pattern, resulting in a representation that is very reminiscent of realistic handwritten instances. The photos created are capable of being used individually to train the device or mixed with natural handwriting data to supplement the initial database to further enhance the method of recognizing letters [6]. A toolbox could be used for creating synthetic Hungarian handwritten text. Ankush Gupta et al. proposed a novel strategy for recognizing text in natural images. The method has two key advantages: it provides a fast and flexible generator for creating synthetic images of text in cluttered scenes, and it integrates synthetic text into background photos in a realistic manner [11]. A Fully Convolutional Regression Network (FCRN) was trained on these synthetic images to efficiently detect text using bounding-box regression across multiple scales and positions. Similarly, Al-Hitawi [17] demonstrated that pre-training on synthetic data followed by fine-tuning on human-annotated data can significantly enhance the performance of language models in text generation.

2.2. Related Datasets for Hungarian HTR.

The main position Hungarian Handwritten Text Recognition (HTR) datasets developed in the previous five years are presented in Table 1. These massive databases, which are frequently found in historical or archival contexts, offer the real deal at the line and word levels. Hungarian portions have been specially included in some bilingual resources. They serve as useful standards for reviewing our suggested HuHTR-Gen toolbox.

Table 1. Hungarian HTR Datasets

Dataset / Resource	Year	Level	Size
T-H-E Dataset (Turkish-Hungarian-English) [18]	2020	Character	~156,000 chars
HOME-Alcar (HOME-Alcar-line) [19]	2021	Line / Medieval	17 manuscripts, line-level
e-NDP (Collaborative edition) [20]	2023	Page / Line	Collection, ground-truth
Hungarian Census Ground-truth [21]	2024	Line / Name fields	Crowdsourced GT (archival)

3. METHOD

Algorithmic strategies are essential in research and academia for extracting first-hand information from raw text corpora. These techniques use qualities and quantitative methodologies, such as photomontage. The utilization of algorithm-based tools creates synthetic data in handwritten domains using the Edouard Belval tool; we leave the machine learning and deep learning approaches for future work [7]. As far as I know, no Hungarian-annotated handwritten middle or huge database is accessible to anyone for collection. So an option was made to expand on a software program already designed for foreign languages and employ it to produce synthetic data for the Hungarian language. So adapting an existing toolkit is a necessity [13]. We collect approximately 200 open-source fonts' Hungarian scripts [8, 9] to demonstrate the synthetic data. Figure 1, shows how we could generate data in efficient way following algorithmic methodology, and illustrates the pipeline of data generation, where Step A applies corpus-based text synthesis and Step B applies augmentation. The three million line-level pairs (picture, text) have been introduced and made accessible to anyone. This might prove significant for training sophisticated network designs and improving outcomes for tasks involving the acquisition and recognition of handwritten words and lines. This middle amount of data could be utilized in the first stage of pre-training vision-language models in a sequence-to-sequence architecture. In addition, it is an easy task to extend the amount of the desired observation for a huge database.

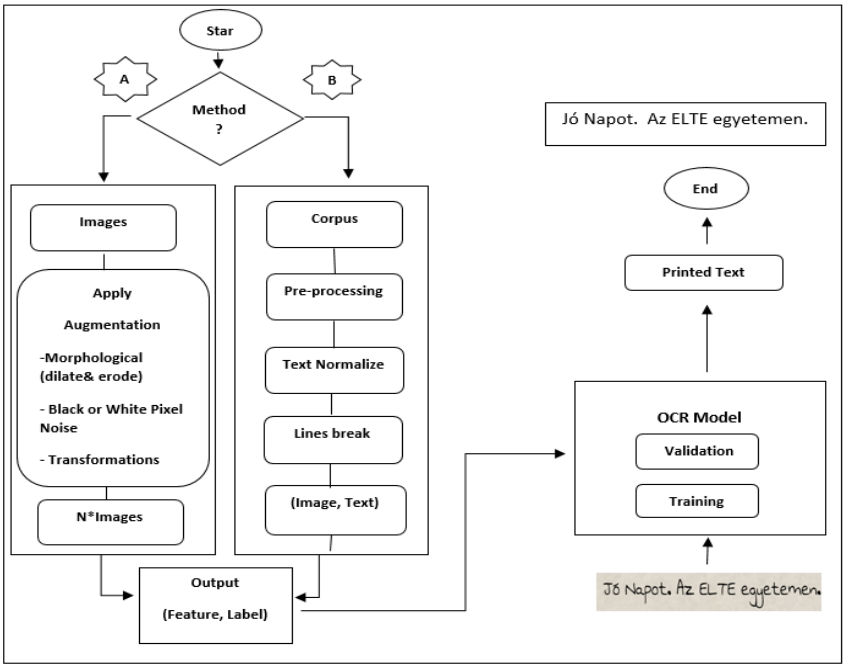


Figure 1. The main pipeline for text recognition data generation. One or both (A, B) of the methods could be used.

4. DATA GENERATION

The scripting method, in particular, where there are many similar traits in huge data sets [10], is crucial in building patterns of recognition tasks. The difficulties that arise include: (i) the probability that a suitable sample allocation can be applied to unknown data, although caution should be exercised due to potential bias. (ii) There is a need for information that is qualitative as well as quantitative.

4.1. Synthetic data

Models are data-intensive, involving billions of attributes. Table 2 shows statistics of the generated samples. The data sources are two Hungarian corpora and three English Brown corpora, with the majority at the line level and a handful at the word level. We sampled this corpus by dividing it into smaller portions, separating long text into predetermined sequences of 8-12 words per line, and sanitizing it by preserving just alphanumeric characters with certain needed particular symbols in this initial stage.

Table 2. Generated synthetic data on lines level

Data	Samples	Language	Level
lines-hu-v1	500000	Hu	line
lines-hu-v2	500000	Hu	
lines-hu-v2-1	935213	Hu	
lines-hu-v3	500000	Hu	
lines-hu-v4	500000	Hu	
lines-hu-v5	500000	Hu	
lines-hu-v6	100000	Hu	
lines-hu-v7	100000	Hu	
Brown	96367	En	
Hu-Words-Dict	60344	Hu	word
Hu-names	4478	Hu	
En-Words-Dict	466479	En	

Then, in the second step, we reconfigure a nice existing toolkit that generates synthetic data by including the Hungarian language. All the needed development steps could be found on the mentioned tool's HuTRDG page. The sources of data are the Hungarian corpus [14] and the English Brown corpus [15].

The Corpus split it into small units and broke long text into fixed sequences of length between 8 and 12 words per line, and cleaned it by keeping only alphanumeric characters with some needed special characters at the first step. In the second step, we adapted an existing synthetic text generator (HuTRDG) to support Hungarian-specific characters and scripts. Data was generated for both English and Hungarian on the word and line level, as shown in Table 1. Seven Hungarian versions are presented; each of these versions (1−7) has different parameters that could be found on the dataset card page for doing augmentation during the generation step, where blur, Gaussian, distortion, rotation, color, uncolored text, noise, and skewing, with different background images. In addition, more parameters can be used. Figure 2 provides random samples of generated lines shows that random samples of generated synthetic data had different backgrounds, and a random font type could be chosen.

Figure 2. Sample of synthetics data generated

4.2. Data Augmentation Performed Efficiently

We use augmentation approaches and use CV methods, which leads to an increase in the meagre dataset. Small adjustments are made every single time a data point goes through training, reducing the likelihood of over-fitting as well as improving generalization. Taking advantage of libraries called Albumentations and OpenCV, as shown in Figure 3, we focus on this area and leave the augmented colour image for further work. i- Morphological alterations have to do with the way the inside is shaped. Putting it simply, we employ this approach to alter the visual appearance of text lines. It is known for expansion and erosion. ii- Noise introduction: Pixels in which colour can be inserted or eliminated. We can use many noise methodologies, but the nearest to converge is to use dark colours with random distribution, so the image will be hard to recognize even for a human. Figure 3 shows augmentation outputs (e.g., erosion, Gaussian noise, skewing).

a- Pixel dropping can be either white or black.

b- Sporadic showers (Add the appearance of rain to the image)

Figure 3. Samples of HTR augmentation method for historical Hungarian data.

The HuHTR-Gen toolkit incorporates a number of scientific components to guarantee uniqueness and realism in the resultant dataset. To investigate handwritten style variability, more than 200 open-source scripts were gathered from platforms like Google scripts and 1001Fonts.

For optimum readability, the language explains how to divide the text through lines of 8–12 words. 256 × 64 pixels were used to make line-level photos and 128 × 32 pixels for word-level images. In addition to morphological steps like erosion and dilation, a variety of augmentation techniques were applied, such as Gaussian blur, salt-and-pepper disturbances, and elastic distortions. Despite font alternation, character spacing jitter, and synthetic stroke deformations boosted the realism of the handwritten text samples, variable backgrounds replicated the real-world circumstances.

5. RESULTS AND DISCUSSION

We used (i) real Hungarian handwriting data (baseline), (ii) synthetic data only, and (iii) synthetic plus real data to generate a transformer-driven HTR model in order to verify HuHTR-Gen. The Hungarian HTR model trained on actual handwritten info obtained a WER of 23.297% and a CER of 5.764%, as shown in Table 3. The area gap was revealed by the considerably worse performance (CER 6.210%, WER 25.814%) obtained through training only on synthetic data. The most impressive outcomes (CER 3.681%, WER 16.091%) were obtained by combining synthetic pre-training with genuine information fine-tuning, demonstrating the improvement in recognition proficiency that synthetic data has to offer. All this results are conducted by fine-tuning the Transformer Optical Character Recognition (TrOCR) introduced by M. Li, et al. Microsoft researcher [22].

Table 3: Experimental results demonstrated as case study based on Transformers architecture.

Training Data	CER (%)	WER (%)
Baseline (real only)	5.764	23.297
Synthetic only	6.210	25.814
Synthetic + real (**ours**)	**3.681**	**16.091**

These findings illustrate that identifying performance is much strengthened by pre-training employing synthetic data and then fine-tuning with genuine specimens.

6. CONCLUSION

This study presents HuHTR-Gen, a data-driven generator and augmentation toolkit for Hungarian handwritten text recognition. The framework has been adapted to the Hungarian alphabet by incorporating additional fonts that support its unique diacritical characters, offering a flexible platform for generating and augmenting datasets at both the line and word levels. To demonstrate its applicability, we employed a transformer-based HTR model as a test case, which confirmed the potential of synthetic data to enhance handwritten text recognition. The generated datasets and toolkit are publicly available, allowing other researchers to integrate them into OCR and HTR tasks, particularly for low-resource languages and cultural heritage digitization.

In future studies, we plan to extend the toolkit to support other low-resource languages with complex scripts, integrate more advanced augmentation methods such as GAN-based handwriting simulation, and conduct broader evaluations using multiple HTR architectures to compare performance across diverse datasets.

ACKNOWLEDGEMENTS

The authors would like to thank Eötvös Loránd University (https://www.elte.hu) in Budapest-Hungary, and University of Fallujah (https://www.uofallujah.edu.iq) in Anbar-Fallujah -Iraq for its support in the present work.

REFERENCES

[1] Kermorvant C 2023 Convergence of OCR and HTR technologies (available at https://teklia.com/blog/202212-atr/).

[2] Sergio Torres Aguilar, and Vincent Jolivet. “Handwritten Text Recognition for Documentary Medieval Manuscripts.” Journal of Data Mining and Digital Humanities, vol. Historical Documents and..., 22 Dec. 2023, https://doi.org/10.46298/jdmdh.10484. Accessed 16 Feb. 2024.

[3] Kiss, Martin, et al. Brno Mobile OCR Dataset. 1 Sept. 2019, https://doi.org/10.1109/icdar.2019.00218. Accessed 26 May 2024.

[4] github.comL. L. János, Á. K., N. L. N., B. N., N. V., G. Z. Y., and T. Váradi, "OCR-hibák javítása neurális technológiák segítségével," in Proc. XVIII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, Jan. 27–28, 2022.

[5] Jha, Ganesh & Cecotti, Hubert. (2020). Data augmentation for handwritten digit recognition using generative adversarial networks. Multimedia Tools and Applications. 79. 10.1007/s11042-020-08883-w.

[6] Roy, Partha & Mohta, Akash & Chaudhuri, Bidyut. (2018). Synthetic data generation for Indic handwritten text recognition. 10.48550/arXiv.1804.06254.

[7] E. Belval, TRDG text recognition data generator. [Online]. Available: https://github.com/Belval/TextRecognitionDataGenerator . Accessed: May 30, 2024.

[8] Google, Google Fonts. [Online]. Available: https://fonts.google.com/?category=Handwriting . Accessed: 2023.

[9] 1001Fonts, *Handwritten Fonts*. [Online]. Available: https://www.1001fonts.com/handwritten-fonts.html. Accessed: 2023.

[10] P. Sankar K, "Nearest neighbor based collection OCR," *Proceedings of the 11th IAPR International Workshop on Document Analysis Systems (DAS)*, 2010, pp. 207–214. doi: 10.1145/1815330.1815357

[11] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 2315–2324.

[12] xReniar, *OCR-dataset-generator* (version latest). [Online]. Available: https://github.com/xReniar/OCR-Dataset-Generator. Accessed: 2023.

[13] Mohammed A.S.Al-Hitawi,OCR_HU_Tra2022:HuTRDG.[Online].Available: https://github.com/Mohammed20201991/OCR_HU_Tra2022/tree/main/HuTRDG. Accessed: 2022.

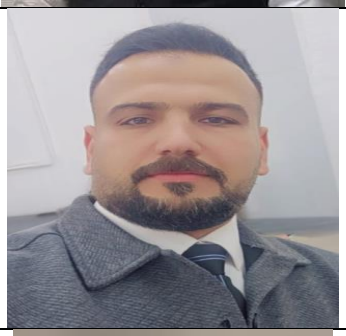
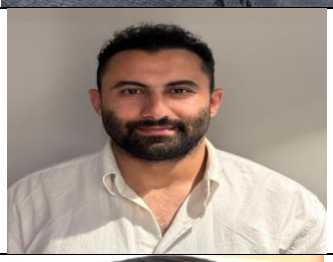
[14] University of Hawai‘i, *Brown Corpus*. Honolulu: Second Language Studies, University of Hawai‘i. [Online]. Available: http://www.sls.hawaii.edu/bley-vroman/brown_corpus.html. Accessed: Aug. 11, 2025.

[15] A. Conneau, G. Wenzek, V. Chaudhary, F. Guzmán, E. Grave, M. Ott, et al., *CC-100: Monolingual Datasets from Web Crawl Data*. [Online]. Available: https://data.statmt.org/cc-100/hu.txt.xz. Accessed: Aug. 11, 2025.

[16] M.A.S. Al-Hitawi, “HuTRDG: Hungarian Text Recognition Dataset Generator (OCR_HU_Tra2022),” 2022. Available: https://github.com/OCR_HU_Tra2022.
[17] Al-Hitawi MSc. Thesis: From English to Hungarian: Fine-tuned OCR. Budapest: Eötvös Loránd University; 2023.
[18] T. Eroglu et al., “T-H-E: Turkish, Hungarian, English Character Dataset for Handwriting Recognition,” 2020.
[19] HOME-Alcar Project, “HOME-Alcar-line Medieval Cartularies Dataset,” 2021.
[20] e-NDP Project, “Collaborative Digital Edition and Ground Truth for Medieval HTR,” 2023.
[21] G. Váradi et al., “Crowdsourced Ground Truth for Hungarian Census Records,” 2024.
[22] M. Li, et al., “Trocr: Transformer-based optical character recognition with pre-trained models,” in Proc. AAAI Conf. Artificial Intell., 2023, pp. 13094–13102.

BIOGRAPHIES OF AUTHORS

The recommended number of authors is at least 2. One of them as a corresponding author.
Please attach clear photo (3x4 cm) and vita. Example of biographies of authors:

	**Lecturer. Mohammed A.S Al-Hitawi**, Received his BSc degree in computer science from University of Anbar –Iraq in 2015 and M.Sc. degree in the artificial intelligence from ELTE University – Hungary in 2023. He has been a full-time lecturer and director of software since March 2019, Computer Center. Currently, he is Artificial Intellgince department rapporteur, at College of Information Technology, University of Fallujah, Iraq, since September 2025, ongoing he can be contacted at email: al_hitawe@uofallujah.edu.iq.
	**Ali L. A. Al-Zaidi** Received his B.Sc. degree in Computer Science from Al-Mustansiriyah University, Iraq, in 2015, and his M.Sc. degree in Information Technology from Altınbaş University, Turkey, in 2020. He is currently serving as a Technical Support Specialist in the Department of Religious Education and Islamic Studies, Iraqi Sunni Affairs, Baghdad, Iraq. His research interests include artificial intelligence, cybersecurity, and machine learning applications. He can be contacted at Ali.layth@taleemdeny.edu.iq .
	**Eng. M A Alazaizi** Received his BSc degree in Civil Engineering from University of Budapest –Hungary in 2023 and. Currently he is M.S.c student at Faculty of Management and Informatics, at Technical University of Munich, ongoing he can be contacted at email:Tech.mohamedahmed@gmail.com.
	**Mustafa Adnan Tharthar** Received his BSc degree in Computer Science from university of Anbar, Currently he is working at the Registration and Students Affairs, University Headquarter, University of Anbar, 31001, Ramadi, Anbar, Iraq and could be contacted at email: Mustafa.alqayssi@uoanbar.edu.iq .