Doaa Sulaiman at AR-MS NakbaNLP 2026: Faithful Diplomatic Transcription of Arabic Manuscripts Using a Human-Centred Annotation Framework

Doaa Bahjat Sulaiman
Institute of Social Sciences, Istanbul University
Türkiye
doaa.sulaiman@ogr.iu.edu.tr

Abstract

This paper describes participation in the Human Transcription Track (Subtask 1) of the NAKBA-NLP 2026 Arabic Manuscript Understanding Shared Task, which focuses on historical handwritten documents related to Palestinian Nakba narratives. The participant manually transcribed 500 cropped line images and designed a comprehensive transcription guideline from scratch, documenting all editorial decisions in a reusable framework. A faithful diplomatic transcription philosophy was adopted, preserving original spelling, punctuation, diacritics, numerals, and layout features without editorial normalisation, to create research-grade gold-standard data for downstream NLP and digital humanities research. Building on this philosophy, a 26-convention annotation framework was developed, organised into three layers: editorial-structural symbols, faithful-copying rules, and documentation labels, supported by a four-step quality-control pipeline. The submission achieved full coverage of all 500 assigned lines and attained an official CER of 0.02, WER of 0.01, and accuracy of 0.98.

Keywords: Arabic manuscripts, handwritten text transcription, Nakba narratives, diplomatic transcription, annotation framework

1. Introduction

Historical Arabic manuscripts constitute an invaluable cultural and linguistic resource, yet their systematic digitisation remains a formidable challenge (Zaghouani et al., 2014; Pierazzo, 2011). The primary obstacles include the pronounced variability of individual handwriting styles, persistent orthographic inconsistencies, the physical degradation of source materials over time, and the absence of agreed-upon transcription conventions (Hamoud et al., 2026; Zaraket et al., 2026). Advances in Handwritten Text Recognition (HTR) and Optical Character Recognition (OCR) have delivered impressive results for Latin scripts; however, Arabic HTR accuracy continues to lag substantially behind, owing to the cursive and context-sensitive character of Arabic writing, the high degree of allographic variation, and the comparative scarcity of well-annotated training corpora (Ahmad et al., 2017; Ahmad et al., 2020; Aabed and Khairaldin, 2024; Al-Mutawa et al., 2024; Waly et al., 2025).

The NAKBA-NLP 2026 Shared Task directly addresses this gap by providing a structured benchmark for both human and automated transcription of manuscript page images drawn from the Omar Al-Saleh Memoir Collection, a corpus of 16 documents spanning 1951–1965 and estimated to contain approximately 6,395 pages and 1.6 million words (Hamoud et al., 2026; Zaraket et al., 2026). Subtask 1 (the Human Transcription Track) requires participants to transcribe approximately 500 cropped line images and to design their own transcription guidelines; evaluation is based on three dimensions: coverage, accuracy (CER and WER), and guidelines quality (Hamoud et al., 2026).

This paper reports on participation in Subtask 1. The principal contributions are: (a) a comprehensive, reusable 26-convention annotation framework for Arabic manuscripts, organised into three functional layers; (b) a four-step quality-control pipeline designed to approximate multi-annotator rigour in a single-annotator setting; and (c) an empirically supported account of how expert human annotation can complement automated HTR pipelines in large-scale manuscript digitisation projects.

2. Related Work

2.1 Arabic Handwritten Text Recognition

Arabic HTR has attracted considerable research interest in the deep-learning era, yet recognition rates remain well below those achieved for Latin scripts (Ahmad et al., 2017, 2020). The KHATT dataset serves as a canonical benchmark; multi-dimensional LSTM-based architectures achieve character recognition rates of approximately 80% on this corpus, while more recent segmentation-free end-to-end models report rates of around 84% (Ahmad et al., 2017, 2020; Aabed and Khairaldin, 2024). Comparative evaluations of four deep-learning architectures (FCN, GFN, VAN, and DAN) on KHATT indicate that the DAN model achieves the lowest character and word error rates, though performance remains substantially below Latin-script benchmarks (Al-Mutawa et al., 2024).

Among transformer-based systems, HATFormer achieves CER values of 8.6% and 4.2% on major public and private Arabic handwriting datasets respectively (HATFormer, 2024). The Qalam model, trained on over 4.5 million manuscript images, attains a WER of 0.80% for handwritten

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 108–112

text recognition tasks (Bhatia et al., 2024). The end-to-end Invizo system reports 0.59% CER on printed Arabic text and 7.91% CER on handwritten Arabic (Waly et al., 2025). Despite these advances, automated systems continue to struggle with historically degraded manuscripts exhibiting irregular layouts, ink bleeding, and idiosyncratic orthographic conventions, reinforcing the need for carefully produced gold-standard transcription data (Hamoud et al., 2026; Zaraket et al., 2026).

2.2 Annotation Guidelines and Gold-Standard Data

The design of annotation guidelines has been recognised as a critical component of any large-scale NLP data-creation effort. High-quality annotated corpora directly determine the ceiling of model performance, and the process of guideline development itself constitutes a scholarly contribution by encoding linguistic and editorial knowledge in a reusable, transferable form (Zaghouani et al., 2014; Pierazzo, 2011). Zaghouani et al. (2014) demonstrate that iterative calibration, inter-annotator agreement measurement, and transparent documentation of editorial decisions are indispensable for reliable annotation.

Digital scholarly editions of Arabic and other historical texts have further proposed stratified transcription models that formally distinguish between faithful data acquisition and interpretive annotation layers (Pierazzo, 2011; TEI Consortium, 2023). Within this tradition, gold-standard datasets are not simply accurate transcriptions, but also carefully documented editorial objects whose conventions can be reused and adapted across projects.

2.3 Diplomatic Transcription

Diplomatic transcription aims to reproduce a source document as faithfully as possible, preserving scribal errors, non-standard orthography, and original punctuation, without introducing editorial corrections or normalisation (Burghart, 2017; Pierazzo, 2011). This approach contrasts with more interventionist editorial models that regularise spelling or silently correct perceived mistakes, thereby obscuring evidence about authors' practices and historical usage.

The TEI framework provides a rich vocabulary for encoding features relevant to diplomatic work, including deletions, additions, unclear text, and gaps (TEI Consortium, 2023). Recent platforms such as eScriptorium build on these principles to support large-scale transcription, layout analysis, and annotation workflows for historical documents (Kiessling et al., 2019).

3. Task Description and Data

3.1 Task Overview

The NAKBA-NLP 2026 Shared Task comprises two subtasks. Subtask 1 (Human Transcription Track) provided each participating team with a batch of approximately 500-line images cropped from historical Arabic handwritten manuscripts. Teams were required to submit both a completed annotations.csv file and a guidelines document covering their transcription conventions, corner-case handling strategies, consistency measures, and ethical considerations.

3.2 Dataset Characteristics

Each batch folder contained PNG line images cropped from full manuscript pages (1953–1965); an annotations.csv template with columns filename, text, source_image, year, page, and line; and full-page contextual images. The manuscripts are drawn from the Omar Al-Saleh Memoir Collection, which documents Palestinian Nakba narratives authored by Omar Al-Saleh. The collection presents diverse palaeographic challenges including cross-period variation in letter forms, historical orthographic conventions diverging from modern standard Arabic, physical deterioration, authorial editorial marks, and the concurrent use of Arabic (Eastern) and Western numeral systems.

3.3 Evaluation Metrics

Submissions were assessed on three criteria: (a) coverage, (b) accuracy, and (c) guidelines quality. Accuracy was measured using Character Error Rate (CER) and Word Error Rate (WER) computed against an expert-verified ground truth.

4. Transcription Methodology

4.1 Transcription Philosophy

A faithful diplomatic transcription approach was adopted in which the original spelling, punctuation, diacritics, and stylistic features of the manuscript are preserved without editorial intervention. No orthographic normalisation was applied; no punctuation was inserted or removed; no grammar corrections were introduced; diacritical marks were preserved where the author placed them and not added where absent; and numeral system choices were retained without conversion.

4.2 Annotation Framework Design

Since no predefined guidelines were supplied by the task organizers, a comprehensive framework was designed from scratch. The framework is organised into three functional layers, reflecting three distinct roles in the transcription process. The framework contains exactly 26 conventions in total: 11 editorial-structural symbols, 12 faithful-copying rules, and 3 documentation label types.

4.2.1 Layer 1: Editorial–Structural Symbols (11 Conventions)

Precise symbols were defined for handling the most common palaeographic challenges. Each

symbol carries an unambiguous operational definition, a specified usage context, and at least one illustrative example drawn from actual manuscript images. Three design principles governed symbol selection: minimal ambiguity, visual distinctiveness, and reversibility.

4.2.2 Layer 2: Faithful-Copying Rules (12 Conventions)

These rules operationalise the diplomatic transcription philosophy at a granular level, governing the handling of orthographic and linguistic features that must be reproduced without modification.

4.2.3 Layer 3: Documentation Labels (3 Types)

To maintain a transparent audit trail, three label types were introduced for use exclusively within the guideline document: Decision, Interpretive Note, and Note. This system ensures that the guideline document functions as a transparent research diary of the transcription process.

4.3 Transcription Workflow

The transcription was completed over approximately three days following a structured, iterative workflow. The familiarisation phase involved examining a sample of full-page and line images to characterise the author's handwriting, identify recurring abbreviations, and map orthographic patterns. Guideline drafting produced an initial convention set focused on the most frequently encountered phenomena, and iterative refinement extended the framework to the final 26 conventions.

5. Quality Control

To ensure high-quality, consistent transcriptions across all 500 lines, a four-step quality-control pipeline was implemented. All transcription work was conducted by a single annotator following the proposed guideline, with multiple self-review passes designed to approximate the rigour of multi-annotator workflows through systematic consistency checks.

Step 1 consisted of periodic batch review. Step 2 consisted of cross-document consistency checks. Step 3 involved anomaly detection for duplicate image IDs, empty text fields, unexpected symbol combinations, and outlier line lengths. Step 4 required transparent documentation of every ambiguous reading decision in the guideline document.

6. Results

The submission achieved full coverage of the assigned batch: all 500-line images were transcribed and included in the final annotations file. The official shared-task evaluation returned a CER of 0.02, WER of 0.01, and accuracy of 0.98, reflecting the high character-level accuracy attained through the combination of diplomatic transcription philosophy and the structured quality-control pipeline.

This result demonstrates that a carefully designed single-annotator workflow, when supported by explicit conventions and iterative review, can produce transcription quality that is competitive with reported automated Arabic HTR systems operating on similarly degraded historical manuscript material.

7. Discussion and Recommendations

The CER of 0.02, WER of 0.01, and accuracy of 0.98 confirm that expert human transcription, guided by a principled convention system, can achieve very low character-level error rates on challenging Arabic manuscript material. The three-layer annotation framework proved particularly valuable in stabilising editorial decisions over time.

A further observation concerns the complementarity between human and automated approaches. Human annotators can efficiently produce small, carefully curated datasets that capture the full palaeographic complexity of a given collection; automated systems subsequently excel at scaling those annotations to tens of thousands of lines. Future shared tasks would benefit from full-page context, explicit scoring criteria for guideline quality, and at least a subset of multi-annotator data.

8. Conclusion

This paper presented a faithful diplomatic transcription approach for historical Arabic manuscripts applied to the NAKBA-NLP 2026 Shared Task. A 26-convention annotation framework organised in three functional layers was developed from scratch and supported by a four-step quality-control pipeline. The resulting submission achieved full coverage of 500 lines with CER of 0.02, WER of 0.01, and accuracy of 0.98.

Limitations

This study was conducted by a single annotator; therefore, formal inter-annotator agreement could not be computed from the submitted data alone. In addition, the 26-convention framework was developed specifically for the Omar Al-Saleh Memoir Collection, and its transferability to other Arabic manuscript collections remains to be tested.

Ethics Statement

The manuscripts transcribed in this task contain historical narratives related to the Palestinian Nakba. This material was approached with respect and cultural sensitivity, maintaining strict fidelity to the original text without editorial

modification. No personal data were collected or processed beyond the materials provided by the task organizers.

Acknowledgements

The author thanks the organizers of the NAKBA-NLP 2026 Shared Task for providing the data, benchmark structure, and evaluation framework that made this study possible.

Bibliographical References

Aabed, S. and Khairaldin, A. (2024). An end-to-end, segmentation-free, Arabic handwritten recognition model on KHATT. arXiv preprint arXiv:2406.15329.

Ahmad, R., Naz, S., Afzal, M. Z., Rashid, S. F., Liwicki, M., and Dengel, A. (2017). KHATT: A deep learning benchmark on Arabic script. In Proceedings of ICDAR, pages 10–14.

Ahmad, R., Naz, S., Afzal, M. Z., Rashid, S. F., Liwicki, M., and Dengel, A. (2020). A deep learning based Arabic script recognition system: Benchmark on KHATT. International Arab Journal of Information Technology, 17(3):299–305.

Al-Mutawa, H. et al. (2024). A comparative study of four handwritten text recognition models in Arabic script. Ingénierie des Systèmes d'Information, 29(6):2243–2250.

Bhatia, G., Nagoudi, E. M. B., Alwajih, F., and Abdul-Mageed, M. (2024). Qalam: A multimodal LLM for Arabic optical character and handwriting recognition. In Proceedings of ArabicNLP 2024, pages 210–224.

Burghart, M. (2017). Transcription or diplomatic edition. In Digital Scholarly Editions. Queen Mary University of London.

Hamoud, H. et al. (2026). NAKBA NLP 2026: Shared task on Arabic handwritten manuscript understanding. In Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with LREC 2026.

HATFormer (2024). Historic handwritten Arabic text line recognition using transformers. arXiv preprint arXiv:2410.02179.

Kiessling, B., Tissot, R., Stokes, P., and Stökl Ben Ezra, D. (2019). eScriptorium: An open-source platform for historical document analysis. In Proceedings of ICDARW, volume 2, pages 19–19.

Pierazzo, E. (2011). A rationale of digital documentary editions. Literary and Linguistic Computing, 26(4):463–477.

TEI Consortium (2023). TEI P5: Guidelines for electronic text encoding and interchange. Available at: https://tei-c.org/guidelines/.

Waly, A. et al. (2025). Invizo: Arabic handwritten document optical character recognition solution. arXiv preprint arXiv:2502.05277.

Zaghouani, W. et al. (2014). Large scale Arabic error annotation: Guidelines and framework. In Proceedings of LREC'14.

Zaraket, F. et al. (2026). AR-MS: Arabic manuscript understanding. In Proceedings of Nakba-NLP 2026, co-located with LREC 2026.

Appendix A: Convention Tables

The appendix below provides the convention tables referenced in Section 4. In the camera-ready version, these materials may be included after the references.

Table 1: Layer 1 : Editorial–Structural Symbols (11 Conventions)

#	Symbol	Name	Usage	Example
1	[...]	Truncation marker	Placed where part of a word is missing or cropped.	1962_p189_l0069.png
2	(?)	Illegible word marker	Replaces a completely unintelligible word.	1956_p054_l0040.png
3	*	Unclear line marker	Marks an entire faded, damaged, or partially illegible line.	1955_p043_l0009.png
4	{...}	Editorial insertion	Encloses text added by transcriber for clarification.	1956_p063_l0031.png
5	|word|	Struck-through correction	Encloses words the author struck through.	1954_p172_l0006.png
6	|	Overlapping-lines separator	Inserted between two lines stacked in one image.	1962_p129_l0034.png

7	~	Added-letter marker	Flags an extra letter the author added.	1953c_p010_l0056.png
8	|.....|	Line-discontinuity marker	Used when segments are not textually connected.	1962_p063_l0069.png
9	« »	French quotation marks	Reproduced as employed by the author.	1959_p052_l0051.png
10	_ _	Parenthetical remarks	Reproduces author's underscore markings.	1964_p011_l0050.png
11	✓	Check mark	Reproduced to preserve author's annotations.	1964b_p069_l0034.png

Table 2: Layer 2 : Faithful-Copying Rules (12 Conventions)

#	Convention	Rule	Illustrative Case
1	Orthographic variations	Preserved exactly. No normalisation.	إلي vs. إلى
2	Original punctuation only	No punctuation added unless present.	Periods/commas as-is
3	Tanwīn notation	Applied strictly per original.	Bare alif retained
4	Numerals	Transcribed in author's system (Arabic/Western).	No conversion
5	Hamzah spelling errors	Preserved as written, even if incorrect.	مسؤول vs. مسئول
6	Tāʾ marbūṭa / hāʾ	Transcribed as they appear.	ه/ ة kept as written
7	Allāh without diacritics	Written without diacritical marks.	الله without shaddah
8	Diacritical marks	Preserved where present; not added.	Tashkīl retained
9	Compound words	Kept as single connected forms.	قائمقام as one unit
10	Incomplete words	Transcribed without supplying letters.	Partial words as-is
11	Semantically unclear words	Transcribed without modification.	No paraphrase
12	Unnecessary prepositions	Extraneous words kept without omission.	Redundant terms retained