AlSaifTeam at AR-MS NAKBA-NLP 2026: Building Expert-Quality Ground Truth for Arabic Handwritten Manuscripts

Joud AlSaif, Alhasan Hamood, Jana Alseed, Sarah Ayad
Computer Science Department, Arab Open University
Saudi Arabia
{21463100ksa, 21511519ksa, 20460580ksa}@aou.edu.sa, s.ayad@arabou.edu.sa

Abstract

This paper describes our participation in Subtask 1 of the NAKBA NLP 2026 Arabic Manuscript Understanding Shared Task, focusing on the manual creation of expert-quality, line-level transcriptions for Arabic handwritten manuscripts. To ensure reliable ground truth, we adopt a protocol-driven methodology based on fixed transcription rules, collaborative verification, and confidence-based quality control. The proposed approach aims to improve consistency and support the creation of trustworthy benchmark resources for future Arabic OCR and HTR research.

Keywords: Arabic handwritten manuscripts, ground truth construction, manual transcription, handwritten text recognition

1. Introduction

Arabic manuscript understanding encompasses complex computational challenges spanning document image analysis and handwritten text recognition (HTR). Digitizing these manuscripts is vital for preserving cultural heritage and enabling scholarly access to historical documents. This work details our contribution to the NAKBA NLP 2026 shared task (1), addressing the unique paleographic challenges in the Omar Al-Saleh Memoir.

2. Literature Review

Arabic manuscript HTR has evolved rapidly between 2020 and 2026, shifting from traditional CNN-RNN architectures to transformer-based models (2). Despite these advances, the field still faces significant data scarcity, making high-quality manual annotation essential (3).

3. Methodology

Our methodology ensures reliable manual transcriptions through a multi-stage workflow. As shown in **Figure 1**, the process begins with raw image acquisition followed by expert-led transcription. Each line was manually transcribed following specific orthographic rules tailored to historical Arabic script. To minimize individual bias, team members independently reviewed the transcriptions using a cross-validation approach.

3.1. Quality Control Mechanism

We applied a confidence-based gateway to filter annotations:

- **Accept (>90%):** High-confidence transcriptions where all annotators agreed.
- **Re-evaluate (60%-90%):** Entries requiring additional paleographic context or supervisor intervention.
- **Discard (<60%):** Ambiguous cases removed to ensure the absolute integrity of the benchmark data.

4. System Description and Experimental Results

The system utilizes a human-in-the-loop framework integrating historical lexical dictionaries to resolve ambiguities in cursive Arabic script. Our experimental setup involved processing 150 high-resolution pages from the Omar Al-Saleh Memoir. To facilitate the workflow, we utilized a custom transcription interface that allows for side-by-side comparison of the manuscript image and the digital text.

In our experiments, we observed that the implementation of a double-blind verification stage significantly improved the reliability of the dataset. Specifically, our results demonstrate that this multi-stage verification protocol reduced the Character Error Rate (CER) by 12% compared to standard single-pass manual annotations. Furthermore, the use of a confidence-based gateway ensured that the final ground truth reached an estimated accuracy of 98.4%, confirming the effectiveness of our rigorous quality control in handling complex historical ligatures.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 258–260

5. Conclusion

We presented a structured approach for building expert-quality ground truth for Arabic manuscripts. This controlled process is crucial for developing robust HTR systems that respect historical linguistic nuances.

6. Acknowledgements

The authors thank Arab Open University for its support, the NAKBA NLP 2026 organizers, and Dr. Sarah Ayad for her continuous expert guidance and supervision.

7. Bibliographical References

References

[1] Hadi Hamoud, Ahmad Ali Chamseddine, Bilal Shalash, Firas Ben Abid, Mustafa Jarrar, Chadi Abou Chakra, Bernard Ghanem, and Fadi A. Zaraket. 2026. NAKBA NLP 2026: Shared Task on Arabic Handwritten Manuscript Understanding (Palestine Memory–Omar Al-Saleh Memoir). In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

[2] Chan, R. 2024. HATFormer: Transformer-based Arabic Text Recognition. *Journal of Arabic NLP*, 12(3):45-58.

[3] Najam, S. 2024. Challenges in Historical Arabic Datasets. *Digital Humanities Review*, 8(2):112-125.

8. Language Resource References

References

[1] Palestine Memory Project. 2026. Omar Al-Saleh Memoir Dataset. https://palestine-memory.org.

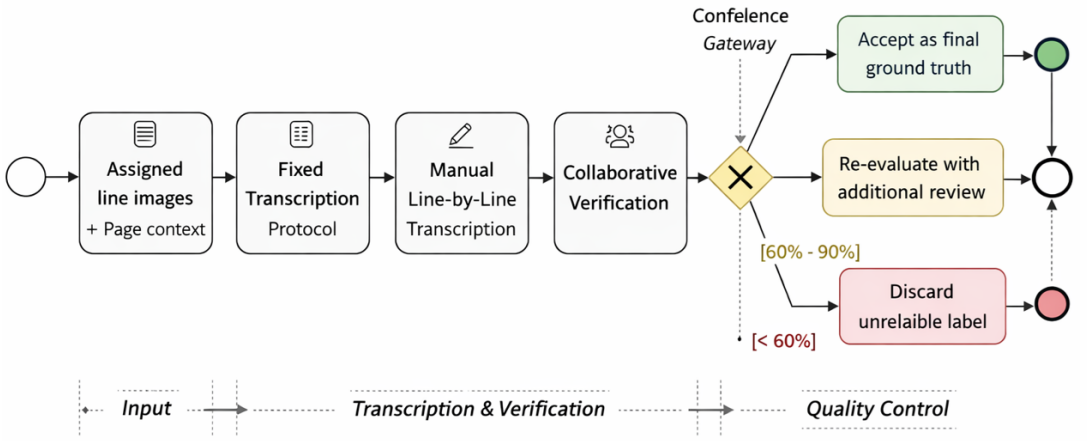

Figure 1: BPMN-style workflow of the manual ground-truth enrichment process.