Baflah-lamri at NAKBA-NLP 2026: Manual Ground Truth Enrichment

Dr. Abdelouahab Bafleh

Scientific and Technical Research Center for the Development of the Arabic Language (CRSTDLA)
Ouargla Research Unit, Algeria

a.baflah@crstdla.dz

Dr. Lamri Mohamed

Scientific and Technical Research Center for the Development of the Arabic Language (CRSTDLA)
Ouargla Research Unit, Algeria

m.lamri@crstdla.dz

Abstract

This paper presents a detailed description of the team's methodology in participating in Subtask 1 (Transcription Track) of the NAKBA NLP 2026 Shared Task for Arabic Manuscript Understanding. We present a rigorous approach to line-level manual transcription of historical Arabic manuscripts derived from the Omar Al-Saleh memoir collection (1951-1965). Our methodology emphasizes accuracy, consistency, and adherence to diplomatic transcription principles, while addressing the unique palaeographic and physical challenges of Arabic handwriting, such as writing speed, orthographic variation, and the impact of writing tools (e.g., immediate strike-throughs and ink spatter). The work guided by strict transcription guidelines and contextual verification protocols, matching cropped line images with full-page images to resolve ambiguities and automated cropping issues. The team successfully transcribed the entire assigned batch of 500 lines across 368 unique pages, producing reference data comprising 6,719 words and 37,646 characters. This effort contributes to providing highly reliable Ground Truth data, serving as an essential foundation for training and evaluating Handwritten Text Recognition (HTR) models for Arabic manuscripts.

Keywords: Handwritten Text Recognition (HTR), Diplomatic Transcription, Historical Arabic Manuscripts, Natural Language Processing (NLP), Ground Truth, Omar Al-Saleh Memoirs, NAKBA NLP Shared Task.

Introduction

The absence of cognitive content from Arabic manuscripts is often due to private ownership and lack of accessibility, particularly for personal memoirs, historical documents, and other archival materials. Given that computational analysis possesses immense capacity and tremendous speed in processing information, making these documents available for computational analysis would be a significant service, enabling researchers to access and benefit from their content.

Automated understanding of Arabic manuscripts represents a critical challenge in the fields of digital humanities and cultural heritage preservation (Chan et al., 2024). Despite its promising potential for preserving cultural heritage and making it widely accessible, it faces numerous challenges primarily stemming from the complexity of recognizing Arabic handwriting, which varies from one scribe to another.

The NAKBA NLP 2026 Shared Task addresses this challenge by providing a benchmark corpus from the Omar Al-Saleh memoir collection, comprising 16 documents spanning from 1951 to 1965, with an estimated 6,395 pages containing approximately 1,597,025 words (Zaraket et al., 2026). This collection exhibits unique characteristics of mid-twentieth-century Arabic handwriting, offering a crucial resource for developing Arabic HTR technologies.

Since computational models cannot decipher Arabic handwriting without a deep understanding of the palaeographic (calligraphic) and linguistic characteristics of the original text, this paper provides a description of the manual transcription methodology applied to a sample of 500 lines from these memoirs, focusing on analyzing the calligraphic and linguistic challenges in the completed batch.

This paper is not limited to presenting quantitative achievements; rather, it offers an analytical reading of the obstacles encountered during the transcription process. It discusses the issues of ambiguous character shapes caused by writing speed and the necessity of relying on context to resolve them. It clearly differentiates between text truncation resulting from automated image cropping and that originating from the manuscript's intrinsic characteristics. Furthermore, it highlights documentary linguistic

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 128–132

phenomena, such as the writing of numerals and the spelling of foreign proper nouns. By presenting these observations on the Ground Truth data, the paper aims to provide qualitative insights that contribute to improving the capacity of deep learning algorithms to process handwritten Arabic texts (with daily memoirs as a case study).

Methodology & Transcription Guidelines

The methodology relied on the principles of Diplomatic Transcription coupled with rigorous contextual verification mechanisms to ensure the construction of highly reliable and historically faithful Ground Truth data. This approach is a fundamental standard in the digitization of historical documents to ensure the text transcribed exactly as written, without linguistic or grammatical corrections that might obscure the original author's stylistic or orthographic characteristics (Zahour et al., 2007).

The transcription process was carried out independently by two native Arabic-speaking annotators to ensure high accuracy and resolve any ambiguities through consensus.

Orthographic Conventions

The transcription process based on a set of strict guidelines to handle recurring orthographic phenomena, which are among the most prominent challenges in Arabic natural language processing (Abandah & Abdel-Karim, 2014):

- **Hamzas:** Strict adherence to the drawing of Hamzas (أ، إ، ؤ، ئ، ء) exactly as they appeared in the original manuscript, regardless of whether they conform to or violate modern standard orthographic rules.
- **Alif Maqsura and Dotted Yaa:** The Alif Maqsura (ى) and the terminal dotted Yaa (ي) transcribed based on their actual visual representation in the text, rather than their linguistic requirements.
- **Ta' Marbuta and Ha':** A precise distinction made between the Ta' Marbuta (ة) and the pronoun Ha' (ه) whenever they were clearly distinguishable. Context used as a deciding factor in cases where dots omitted due to rapid writing.
- **Diacritics:** Transcription was limited exclusively to the diacritical marks clearly visible in the manuscript, ignoring any attempt to infer or add invisible diacritics.

Handling Special Cases

Based on the nature of the images provided in this batch, and guided by the task's instructions (Zaraket et al., 2026), the methodology necessitated the following rules:

- **Unreadable Words and Handwriting Complexity:** Although contextual verification successfully resolved most ambiguities, approximately 20 lines (about 4%) in the completed batch contained entirely incomprehensible words (averaging one to three words per line). This obscurity is not due to physical damage or smudging, but rather to the severe complexity and reduction in character drawing resulting from rapid writing.
- **Marking Missing Words:** Sequential dots (e.g., ...) were used as a placeholder instead of complex tags like [###]. The length of the dot sequence reflects the estimated size of the unreadable word(s) (as seen in reference lines 34, 38, 42, and 46 of the dataset), to accurately locate semantic interruptions.
- **Margins and Insertions:** Neither significant marginal notes nor insertions were recorded in the transcribed batch. In the rare instances of their appearance, the focus remained on the main body of the line (e.g., Line 61).
- **Numerals:** Numbers transcribed in their original format, whether they took the Eastern Arabic form (٠-٩) or the Western form (0-9), reflecting the author's style in recording dates and calculations. In one instance, "962" used, referring to 1962.
- **Foreign Words:** Foreign proper nouns or non-Arabic terms were transcribed exactly as the author wrote them using Arabic letters (phonetic Arabization), preserving Latin characters whenever they were written in the original document.

Dataset Statistics

As a culmination of the adopted "Diplomatic Transcription" methodology, the transcription process resulted in the construction of a reference dataset that faithfully reflects the content of the original manuscript. This section presents a quantitative analysis of the results achieved on the assigned batch, reviewing the coverage scope, statistical metrics of the extracted texts, and readability quality indicators.

Transcription Statistics

The table below reflects the quantitative characteristics of the transcribed data, illustrating the readability level and the volume of extracted texts:

Category	Count / Percentage
Fully readable lines (no missing words)	451 (90.2%)

Category	Count / Percentage
Lines containing unreadable words (marked with ...)	49 (9.8%)
Lines with uncertainty markers or damaged text (e.g., [?])	0 (0.0%)
Lines with foreign language content (Latin characters)	5 (1.0%)
Average characters per line	75.3 ± 16.7
Average words per line	13.4 ± 3.2

Table 1: Transcribed Data Characteristics

Quality Metrics and Linguistic Characteristics

Based on internal verification and statistical analysis of the final transcription file, the following metrics achieved:

Readability Integrity: The percentage of lines that we were able to read and transcribe fully reached 90.2%, while lines containing unreadable words (due to speed and complexity) accounted for 9.8%.

Textual Output: The transcription produced 6,719 words, comprising 37,646 characters.

Numerical Diversity: The memoirs featured a notable presence of numerical data (dates, calculations), with 113 lines (22.6%) containing numbers.

Linguistic Diversity: Foreign names or terms written in Latin characters appeared in five lines (1.0%), reflecting the documentary and political nature of the memoirs.

Challenges & Observations

The transcription process encountered various palaeographic and linguistic challenges. This section presents the most prominent difficult cases observed and how they addressed systematically to ensure data accuracy:

Ambiguous Character Shapes and Handwriting Complexity

The memoirs written in a precise Mashriqi Naskh script, reflecting the author's personality. He adhered to distinctive shapes for certain letters, such as the terminal Nun (written above the line) and the Ha' preceding a final Alif. There was also a special ligature for juxtaposed Ba' and Ta'. Familiarity with these shapes eliminated reading difficulties. However, other challenges arose:

Challenge: Visual confusion of structurally similar letters (especially distinguishing between Ba', Ta', Tha', or terminal Nun and Yaa) when diacritical dots are faded due to rapid writing.

Solution: Examining the surrounding context to verify lexical plausibility and continuously referring to the full-page image to identify the author's usual pattern. In extreme cases where the word's shape was entirely reduced (49 lines), guessing was avoided, and the word was marked with sequential dots (...) to preserve scientific integrity (e.g., Lines 21 and 23).

Foreign Names and Orthographic Variation

Challenge 1 (Foreign Names): Transcribing the names of foreign figures (American and European) written in Arabic letters without certainty of their standard spelling.

Solution: Transcribing the text exactly as written, preserving the original phonetic Arabization used by the author, without attempting to "correct" it (e.g., keeping "فيليب هوف" [Philip Hoff] exactly as the author wrote it).

Challenge 2 (Orthographic Variation): The inconsistent use of certain orthographic rules within the same document, specifically regarding Hamza placement (e.g., مسألة/مسءلة, مبدأ/مبدا).

Solution: Strictly adopting Diplomatic Transcription by transferring the text as is without linguistic intervention, which documents historical variation patterns for future analysis.

Automated Cropping Challenges and Extended Sentences

Challenge 1 (Extended Sentences): Some cropped lines contained a truncated part of a long sentence, making the semantic meaning and syntactic context incomplete.

Solution: Utilizing the full source page image (source_image) to support understanding the overall context of the paragraph, enabling more accurate transcription of the truncated words.

Challenge 2 (Image Cropping): The appearance of horizontal or vertical truncation of letters at the beginnings or ends of lines (e.g., Line 23).

Solution: Methodologically distinguishing between document defects and processing defects. By returning to the original page, it confirmed that this truncation resulted exclusively from the automated cropping process and was not a physical omission.

Physical Constraints & Writing Tools

Challenge 1 (Ink Spatter and Pen Pressure): The use of multiple pens was noted, manifesting in scattered ink dots. Variations in pen pressure produced multiple visual representations for the same letter.

Solution: Overcoming this by zooming in and utilizing context, alongside referencing the original page to verify whether a spot was a diacritic or a transient ink mark.

Challenge 2 (Immediate Strike-throughs): Some lines included rapid strike-throughs (crossing out a word and writing another above or next to it), creating visual noise.

Solution: In adherence to diplomatic transcription, crossed-out words ignored. Focus exclusively placed on transcribing the approved, final text that forms the correct context, to prioritize the extraction of the author's final intended text, avoiding complex tags that might confuse training models.

Challenge 3 (Printed Text): The presence of printed text within the handwritten text, likely because the paper belonged to pre-printed notebooks.

Solution: Adopting only the handwritten text and ignoring the printed text, as it falls outside the context.

Discussion

The work accomplished in this batch offers a qualitative contribution to enriching the Ground Truth data for developing Arabic HTR models. Analyzing the challenges yields several methodological implications:

Palaeographic Characteristics of the Memoir Style

The transcription process demonstrated that "personal memoirs" pose challenges that do not fundamentally differ from carefully copied official documents. The 9.8% of lines containing unreadable words is not a flaw in the document or scanning quality; rather, it reflects an emotional state or rapid flow of thought that led to character reduction. Furthermore, the impact of writing tools (varying pen pressure, ink spatter, and strike-throughs) adds a dynamic character. This provides a crucial indicator: future algorithms must trained on Language Models to predict words based on context, rather than relying merely on the raw visual image.

Impact of Digital Processing (Automated Cropping)

The phenomenon of "truncation" in extended letters—resulting from automated cropping and necessitating continuous reference to full-page images—highlighted a technical weakness in data preparation. This confirms that developing Line Segmentation algorithms tailored to the nature of the Arabic script (characterized by vertical overlap and sub-line extensions) is a fundamental prerequisite for successful automated recognition.

Linguistic Variation and the Importance of Flexible Lexicons

Observations regarding the phonetic Arabization of foreign names and orthographic variation reflect the absence of strict standardization in daily Arabic writing during the mid-twentieth century. This lexical and orthographic diversity requires computational model designers to build Flexible Lexicons capable of accommodating spelling multiplicity without classifying them as "errors".

Conclusion

This paper presented an analytical description of our team's methodology in completing Subtask 1 of the Arabic Manuscript Understanding Shared Task (NAKBA NLP 2026). By applying the principles of Diplomatic Transcription and relying on contextual verification, we successfully transcribed 500 lines assigned rows 100 %. As noted in our challenges, approximately 20 lines contained between one to three unreadable words due to handwriting complexity. To build a highly reliable reference dataset comprising 6,719 words representing historical memoirs spanning between 1951 and 1965.

Our experience demonstrated that the core challenges in dealing with personal memoirs do not lie in physical damage, but rather stem from the characteristics of rapid handwriting (strike-throughs, character reduction, and varying pen pressure), alongside the technical challenges resulting from automated cropping. This batch also provided insights into orthographic variation and the Arabization of foreign names.

Finally, in affirmation of open science principles, all transcribed texts will made available via a dedicated Anonymous GitHub repository under a Creative Commons license (CC-BY-4.0). We hope that this data and the accompanying methodological analyses will contribute to guiding and developing Arabic HTR systems, and to preserving and analyzing Arab documentary heritage more effectively.

Bibliographical References

Abandah, G. A., & Abdel-Karim, S. (2014). Survey and bibliography of Arabic optical text recognition. *Signal Processing*, *98*, 447–451. https://doi.org/10.1016/j.sigpro.2013.11.026

Al-Shatnawi, A., & Al-Zoubi, M. (2024). A systematic literature review of deep learning methods for Arabic handwritten text recognition. *Engineering, Technology & Applied Science Research*, *14*(4), 15234–15245. https://doi.org/10.48084/etasr.6913

Chan, A., Mijar, A., Saeed, M., Wong, C.-W., & Khater, A. (2024). HATFormer: Historic handwritten Arabic text recognition with

Transformers. *arXiv*. https://doi.org/10.48550/arXiv.2410.02179

Zahour, A., Likforman-Sulem, L., Boussellaa, W., & Taconet, B. (2007). Text line segmentation of historical Arabic documents. In *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)* (pp. 138–142). IEEE. https://doi.org/10.1109/ICDAR.2007.4378688

Zaraket, F., Shalash, B., Hamoud, H., Chamseddine, A., Abid, F. B., Jarrar, M., Chakra, C. A., & Ghanem, B. (2026). AR-MS: Arabic manuscript understanding. 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, May.*