PalNLP at AR-MS Shared Task: Guidelines Paper for Arabic Manuscript Transcription

Mutaz Ayesh
Cardiff University
AyeshMA@cardiff.ac.uk

Abstract

This paper describes the guidelines that the PalNLP team recursively developed and followed during the transcription of the assigned batch, as part of the AR-MS Shared Task. The team, which consists of a single experienced transcriber, has manually transcribed 500 images of lines from the Omar Al-Saleh Memoir Collection.

Keywords: transcription guidelines, OCR, Arabic transcription

1. Introduction

This document presents the transcription guidelines that were synthesized and followed as part of the Arabic Manuscript Understanding Shared Task organized by Nakba-NLP 2026 (Hamoud et al., 2026). The team, which consists of a single experienced annotator, received a subset of the Omar Al-Saleh Memoir Collection to transcribe. The aim of the task was to support Arabic optical character recognition (OCR), and specifically handwritten text recognition (HTR).

The subset contained 500 images of lines, spanning 13 years of mostly handwritten material (1953-1965). Few sentences from the subset were typed. Images were line-level crops, not full pages. The team, however, also received the full context of these lines, which proved useful.

The methodology can be described as one of strict alignment with the handwritten lines, especially on the character level. Orthographic details were preserved and normalization or editorial or linguistic corrections were avoided whenever possible. This strict alignment also encompassed preserving incomplete or truncated words at the beginning or end of the line-based images.

Metric	Value
Total Lines / Segments	500
Total Characters	38,252
Total Words	6,757
Vocabulary Size (Unique Words)	4,118
Average Words per Line	13.51
Edge Cases	55

Table 1: Summary statistics of the batch.

2. Guidelines and Examples

There were four main guidelines that guided the transcription process, which were provided by the organizers: (1) transcribing the text exactly as it appears; (2) preserving the original spelling and character forms of the author; (3) preserving punctuation marks, numbers, and diacritics in their original forms; and (4) providing the best interpretation for unclear, damaged, or illegible text. Multiple instructions were extracted accordingly.

Footnotes indicate the identifiers corresponding to each line crop, usually formatted as year_page_line.

2.1. Arabic-Indic Digits

Arabic-Indic digits were preserved as in the source images, with no normalization to Western numerals (0–9), as shown in Table 2.

Transcribed line	Notes
١٤ صفر[1]	"14 Safar".
١٣٨٢ هجرية[2]	"1382 Hijri".

Table 2: Examples illustrating the preservation of Arabic-Indic digits in the transcription.

2.2. Typos

Table 3 provides examples of writing errors that were preserved in the transcription.

Transcribed line	Notes
(٣) دارسة النواحى العسكرية في اللجان المختصة ونضع مخططا يستند على هذه الدعائم[3]	دارسة instead of دراسة due to the *alef* clearly preceding the *ra'*.
الحرآسة الليلية والنهارية مدر العام ولا يسمح لهؤلأ الافراد المدربين ان يغادروا البلاد الا[4]	The unnecessary *madda* diacritic in الحرآسة; مدر instead if مدار; and هؤلأ instead of هؤلاء.

Table 3: Instances of uncorrected typos.

[1] ID: 732294a1c8ef4d55ad023c877c098e26-0004-37
[2] ID: 90aa2437fec34a18bf1fad0830294b33-0017-08
[3] ID: 1962_p086_l0022
[4] ID: 1959_p134_l0017

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 261–264

2.3. Incomplete Words

Incomplete words, whether in the beginning or end of the line were transcribed. When necessary, the *tatwīl* character was used to ensure that the truncated word matches the handwritten text, as shown in Table 4.

Transcribed line	Notes
قرأت الجرائد المحلية وتتبعت الاحداث والحوادث تستغرب ولا تكاد تصدق فر[5]	The incomplete word here is فر at the end of the line.
انني اكره التدخين وارى فيه الضرر وانما اخفف واقلل واتيه ع[6]	The *tatwīl* character was used due to the word-initial *ayn*.

Table 4: Examples showing the transcription of incomplete words.

2.4. Preservation of Original Orthography

The letter ى was transcribed in its original form even in cases where the standard orthography would require the dotted counterpart ي. This includes word-final instances typically written with ي, such as لأنني, شؤوني, and في. Two representative examples are shown in Table 5.

Transcribed line	Notes
احضر الى ابن سعود العشوة صندوق سيجار وذهبت ليلا الى الكعيبني لقضاء مصلحة ابو كحيل واحضر لي مشهور[7]	الى is usually spelled as إلي.
وزارة لبنان: اشيع ان رشيد كارمي استقال بعد مساع كثيرة ولا يهمنى هذا النبأ[8]	The standard spelling of يهمنى is يهمني.

Table 5: Orthographic adherance to the handwritten text, as exemplified by the ى.

2.5. Whitespaces

Whitespace was preserved exactly as it appeared in the source images. No normalization or correction of spacing was applied, including instances where spaces occurred between words and punctuation marks. For example, the author occasionally inserted spaces before punctuation, which were retained in the transcription, as shown in Table 6. The second example contains extended spacing between two verses of a poem, which is a common formatting convention in Arabic poetry.

[5]ID: 1959_p151_l0015
[6]ID: 1962_p045_l0026
[7]ID: 1959_p158_l0079
[8]ID: 1958_p123_l0021

Transcribed line	Notes
اتنى هدية ازهار ملأت اماكنها واغدقت على النفس بهجة وراحة .[9]	A whitespace was inserted between the word and the full-stop to reflect the written text.
كما قالت تلك الشاعرة: عجبا عجبا له ابكيه ملأ مدامعي وأقول لا شلت يمين القاتل.[10]	Multiple white spaces were inserted in our transcription to reflect those found between the verses.

Table 6: Examples illustrating the preservation of whitespace in the transcription.

2.6. Diacritics

Diacritics were preserved exactly as they appeared in the source images, including marks such as *madda* and *shadda*. No normalization was applied to *hamzat* forms (e.g., أ, إ, ؤ, ئ) even in cases where their insertion or modification might have been grammatically expected, ensuring that the transcription reflects the text as written rather than a corrected version.

Transcribed line	Notes
ينظمون صفوفهم ويهيؤون رجالهم لحرب صاحقة اخشى ان تؤدي العقبي لضياع البقية[11]	The *hamza* was preserved due to appearing in the original image.
ضباط الجيش سموا انفسهم "الضباط الاحرار" واصدروا نشرات وزعوها في اجواء الجيش[12]	The *hamza* was not introduced in any of the underlined words where it would be expected (انفسهم, اجواء, اصدروا, الاحرار).
عدوا منافسًا الى السّعودية في الخليج؟ انى ارى هذا من السياسة الاستعمارية القديمة[13]	This example shows a *shadda* and a *tanwīn*.

Table 7: Instances that required a Google search. The side search helped disambiguate the spelling.

3. Challenging Cases

3.1. Instances that Required a Side Search

Google Search was utilized exclusively for ambiguous handwritten text involving foreign words or named entities (see Table 8). To ensure the transcription remained faithful and free from external bias, search results were used solely to generate spelling "hypotheses" for unclear strokes, which

[9]ID: 1958_p087_l0005
[10]ID: 1962_p177_l0049
[11]ID: 1955_p184_l0035
[12]ID: 1955_p176_l0036
[13]ID: 1956_p047_l0027

were then cross-referenced against the manuscript. A term was only finalized in the transcription if the visual evidence completely supported the external Google results. This ensured that the document itself dictated the final text rather than outside knowledge.

Transcribed line	Notes
اتى هزاع وزير الداخلية القائم الى رام الله فالتف حوله ارباب المصالح ودعاه حنا العسعس.[14]	هزاع and حنا العسعس are Palestinian figures.
حزب آخر الى مساعدته وقد سقط اساطين اعداء ديغول امثال رينو ومنديس فرانس وغي موليه وغيرهم[15]	منديس فرانس, رينو, ديغول, and غي موليه are French politicians.

Table 8: Instances that required a Google search. The side search helped disambiguate the spelling.

3.2. Incomprehensible Text

When a segment of text was difficult to decipher, it was transcribed exactly as it appeared in the line image without attempting correction or normalization. In such cases, the surrounding page context (provided by the organizers) was used to aid interpretation. While contextual information helped resolve ambiguous words (Example 2 in Table 9), it did not clarify all instances (Example 1).

Transcribed line	Notes
اذا حز به الاغياء وهزه ما زعزع مجرى حياته العادية انقلب كل ما فيه من مرح الى ترح[16]	The second and fourth words are incomprehensible.
حكما لا وزن له ولا اهمية. كم اصبر وكم اتفاضى ولكن الصبر انطوى واليد قصيرة والناس هائمون[17]	The words كم اصبر وكم looked ambiguous due to the lack of space between them in the source image. This required reading the lines before and after it to decipher the handwritten text.

Table 9: Examples of incomprehensible handwriting that was transcribed to the closest visually identifiable reading.

To maintain a transparent uncertainty policy, any text that remained incomprehensible was transcribed to its closest visually identifiable reading and explicitly underlined and annotated with an "incomprehensible" comment. This approach ensures that future readers can easily distinguish between confident transcriptions and visual approximations. PalNLP would be delighted to share the annotated spreadsheet with the organizers.

3.3. Multiple Lines in one Image

In a small number of cases, the provided line images contained text written across more than one visual line. When this occurred, the transcription preserved the original structure by inserting a newline to reflect the layout of the source image. In addition, the author occasionally positioned numbers (e.g., dates or times) beneath the noun they modify. In such cases, the numbers were transcribed inline immediately after the relevant word. Examples of these cases are shown in Table 10.

Transcribed line	Notes
جثمت في الفراش انتجاعا للراحة وحبا في الابتعاد عن الناس وهذه خير حكمة ادرعها. دفعت حساب الكهرباء ثلاثة دنانير و٣٧٠ فلسا.[18]	Line break preserved.
كل مصيبة اليمة. ففي دقيقة واحدة اختفى عدة الوف من الايرانيين. عند الساعة ٩,٢٠ من ليل السبت[19]	The number ٩,٢٠ was transcribed right after the word "الساعة" despite appearing below it in the source image.

Table 10: Examples of multi-line text.

4. Conclusion

This brief guidelines document aimed to bring context and structure to the PalNLP team's submission for Subtask 1 of the Arabic Manuscript Understanding Shared Task. The transcription process generally aimed to reflect the handwritten text as much as possible so that OCR and HTR systems trained or evaluated on this batch can be assessed under realistic conditions that preserve orthographic variation, diacritics, ligatures, and other manuscript-specific elements.

Although the transcription was produced by a single annotator, consistency was supported through iterative self-review, revisiting earlier decisions when recurring handwriting patterns became clear, marking uncertain segments for later verification, and repeatedly checking batch-level agreement on CodaBench against the expert-verified reference.

[14] ID: 1955_p110_l0009
[15] ID: 1962_p175_l0019
[16] ID: 1959_p107_l0061
[17] ID: 1956_p127_l0051
[18] ID: 1958_p139_l0028
[19] ID: 1962_p136_l0010

References

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.