

People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
Higher School of Computer Science
8 Mai 1945 - Sidi Bel Abbès



Engineering Thesis

To obtain the diploma of **Engineering Degree**

Field of Study: **Computer Science**

Specialization: **Artificial Intelligence & Data Science (IASD)**

Theme

**AraMS-28k: A Weakly Supervised Pipeline for
Line-Level, Layout-Aware Annotation and Restoration
of Historical Arabic Manuscripts**

Presented by
Zellagui Mohamed Diaa Eddine
Guechaoui Mohamed

Defended on: **July 2026**
In front of the jury composed of

Dr. KHALDI Belkacem
Dr. Chaib Souleyman
Dr. BOUDIEB Wassim

President of the Jury
Thesis Supervisor
Examiner

Abstract

This thesis presents three integrated contributions to the computational processing of historical Arabic manuscripts: a weakly supervised annotation pipeline, the AraMS-28k dataset, and a manuscript restoration subsystem.

First, the thesis develops a weakly supervised annotation pipeline that combines page segmentation, multimodal large language model (LLM) structured OCR, and fuzzy text alignment to automate line-level annotation generation while preserving complete human oversight. A central methodological contribution is the *confidence-100 rule*: we demonstrate both theoretically and empirically that when the normalized fuzzy similarity between an OCR line and its matched ground-truth span reaches its maximum value, the correspondence is correct under the adopted matching assumptions—a property we justify formally and then verify empirically—notwithstanding differences in diacritization. Manual review of all confidence-100 samples across 14 books yielded zero observed matching errors, confirming the reliability of the validation protocol.

Second, using the developed pipeline, we construct and release AraMS-28k, a line-level annotated historical Arabic manuscript dataset that is, to our knowledge, the first to provide explicit marginal-annotation labels at this scale, and whose multi-script coverage (including Maghrebi and Ruq’ah hands) is not matched by existing public resources. It comprises 14 books, 548 fully annotated pages, 2,495 partially annotated pages, 27,969 main-text line samples, and 131 marginal line samples. The dataset includes bounding boxes, main/margin labels, insertion anchors for margin lines, confidence scores, and full provenance information. The framework achieves a $6.5\times$ effective speed-up over manual transcription workflows. To establish the dataset as a usable benchmark, we train a Transformer line recogniser on it and report a stratified cross-script evaluation: 6.48% character error rate on in-distribution pages, rising to 25.37% on a fully unseen Naskh book and 37.88% on unseen Maghrebi script — a gradient that quantifies, rather than merely asserts, the generalisation gap the dataset is built to expose.

Third, the thesis introduces a manuscript restoration subsystem comprising a synthetic degradation model applied on the fly during training (four freshly-randomised degraded variants per clean crop, resampled every epoch) and a U-Net restoration network enhanced with a *text-aware* training objective: a frozen HATFormer recogniser reads the restored image and its recognition error against the ground-truth transcription is back-propagated into the restorer. Following the recognition-based evaluation of AutoHDR, the restorer is assessed on both image-quality metrics and the downstream OCR error of the restored output; on the held-out test set the text-aware objective improves downstream recognition over a recognition-agnostic U-Net baseline at near-equal structural fidelity (SSIM 0.978

versus 0.989), reducing character error rate from 35.1% on the degraded input to 27.1% — against 28.1% for the recognition-agnostic baseline — and recovering the corrupted text to within about one point of the 26% clean-line recognition level.

Keywords: Arabic manuscript OCR; weak supervision; LLM-assisted annotation; fuzzy string alignment; document layout analysis; historical document processing; manuscript restoration; marginal annotation.

Resume

Ce memoire presente trois contributions integrees pour le traitement informatique des manuscrits arabes historiques: un pipeline d’annotation faiblement supervise, l’ensemble de donnees AraMS-28k, et un sous-systeme de restauration de manuscrits.

Premierement, le memoire developpe un pipeline d’annotation faiblement supervise combinant la segmentation de pages, l’OCR structuree par grand modele de langage multi-modal (LLM), et un alignement flou de texte pour automatiser la generation d’annotations au niveau ligne tout en preservant une supervision humaine complete. Notre contribution methodologique centrale est la *regle de confiance-100*: nous demontrons formellement et empiriquement que lorsque la similarite floue normalisee atteint sa valeur maximale, la correspondance de ligne est correcte, independamment des differences de vocalisation. L’accord inter-annotateurs ($\kappa = 0,94$) confirme la fiabilite du protocole de validation.

Deuxiemement, a l’aide du pipeline developpe, nous construisons et diffusons AraMS-28k: 14 ouvrages, 548 pages entierement annotees, 2,495 pages partiellement annotees, 27,969 echantillons de lignes de texte principal et 131 echantillons de lignes marginales. Le pipeline atteint une acceleration effective de $6,5\times$ par rapport a la transcription manuelle.

Troisiemement, le memoire presente un sous-systeme de restauration comprenant un modele de degradation synthetique applique a la volee pendant l’entrainement (quatre variantes degradees re-tirees aleatoirement par imagerie, re-echantillonnees a chaque epoque) et un reseau U-Net enrichi d’un objectif *sensible au texte*: un reconnaisseur HATFormer fige lit l’image restauree et son erreur de reconnaissance par rapport a la transcription de reference est retropropagee dans le restaurateur. Suivant le protocole d’evaluation par reconnaissance d’AutoHDR, le restaurateur est evalue a la fois sur des metriques de qualite d’image et sur l’erreur OCR en aval; l’objectif sensible au texte ameliore la reconnaissance par rapport a une base U-Net non sensible au texte, a fidelite structurelle quasi identique (SSIM 0,978 contre 0,989), reduisant le taux d’erreur de caracteres sur l’ensemble de test de 35,1% (entree degradee) a 27,1% (contre 28,1% pour la base non sensible au texte).

Mots-cles: OCR de manuscrits arabes; supervision faible; annotation assistee par LLM; alignement flou de chaines; analyse de la mise en page; restauration d’images de documents.