Transkribus User Conference (TUC) 2026
21 – 23 September 2026, University of Passau

Public Syriac HTR Models on Transkribus: Benchmarks and Guidelines for Multi Scripts

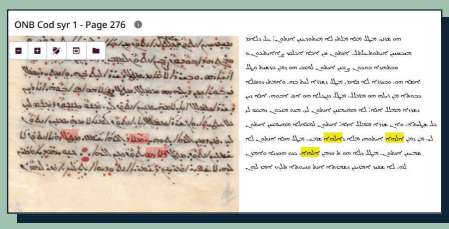
In the framework of the FWF Projects: "Identifying Scattered Puzzles of Syriac Liturgy Manuscripts" (**DOI:** 10.55776/P36649) & the Cluster of Excellence "EurAsian Transformations" (**DOI:** 10.55776/COE8).
Presented by: Ephrem Aboud Ishac, PhD (aboud.ishac@oeaw.ac.at).
Research Institution: Institute for Medieval Research, Division of Byzantine Research, Austrian Academy of Sciences, Vienna.

I. The Philological and Infrastructural Challenge

THE CONTEXT:
The Syriac liturgical heritage exists as a globally dispersed, highly fragmented corpus. Handwritten Text Recognition (HTR) serves as an urgent reconstructive methodology to reunite these scattered archives virtually.

THE RTL BREAKTHROUGH:
Early digital humanities platforms were algorithmically optimized for Latinate, left-to-right spatial logic. By identifying and resolving a fundamental Right-to-Left (RTL) directional bias within Transkribus, this initiative unlocked robust HTR capabilities for the entire RTL ecosystem, including Arabic and Hebrew.

THE DEMOCRATIZATION IMPERATIVE:
True digital democratization requires lowering technical barriers. By removing the need for advanced coding skills or expensive, self-hosted server infrastructure, these public models empower a diverse "digital public"—from independent researchers to cultural heritage practitioners—to access advanced HTR.



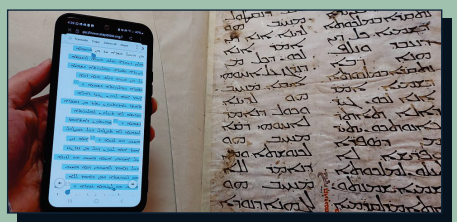
II. Establishing Ground Truth: The "Major Editor" Framework

Syriac scripts feature a complex system of diacritics where a single dot dictates semantic meaning (e.g., distinguishing the active participle *omar* from the perfect tense *emar*). To prevent algorithmic "philological chaos," we implemented rigorous ***Syriac HTR Guidelines*** enforced by a designated "Major Editor":

PHILOLOGICAL TAXONOMY:
Morphological dots (such as plural markers and feminine suffixes) are strictly preserved, while phonetic vowel dots (e.g., ***qushoyo, rukokho,*** and ***linea occultans***) are systematically excluded to ensure machine consistency.

SPATIAL ARCHITECTURE:
Algorithmic precision is secured through standardized regional tagging (MainText, MarginalText) and meticulous manual baseline adjustments.

STRUCTURAL ONTOLOGY:
Scribal loyalty is maintained by transcribing historical punctuation exactly as it appears in the manuscript, avoiding modern interpolations.

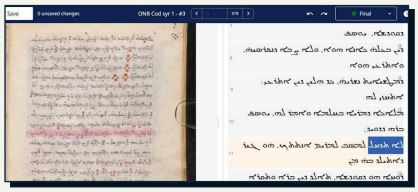


III. State-of-the-Art Public Benchmark Models

This initiative pioneers the digital reconstruction of the Middle Eastern manuscript tradition by releasing the first publicly available, multi-script models on Transkribus.

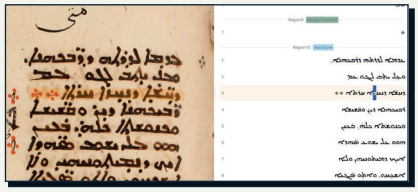
1. SERTO: THE VIENNA SYRIAC GOSPELS

- **SOURCE:** An exceptionally small (91 x 75 mm), text-dense manuscript copied by the renowned scribe Moses of Mardin in 1554 - Vienna.
- **Result:** Established an accurate baseline with a 5.97% Character Error Rate (CER) (Model ID 267309) based on 125 training pages and 20,381 words.
- **Digital Resource:** A fully searchable public website based on the model's training dataset is available online *(Vienna Syriac Gospels - Moses of Mardin,1554).*



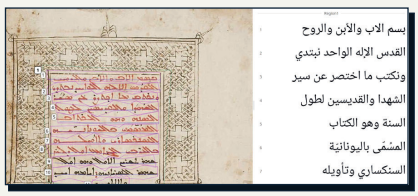
2. ESTRANGELO-MELKITE: MS 36 (JERUSALEM)

- **SOURCE:** A 12th–14th century New Testament featuring complex ligatures, low line distance, and fluid script variations by the scribe.
- **RESULT:** Achieved a validated, state-of-the-art CER of 3.53% (Model ID 468925) utilizing a 56,214-word Ground Truth integrated with the SegmOnto structural vocabulary.



3. GARSHUNI-ARABIC (SYRIAC-ARABIC ALLOGRAPHY)

- **SOURCE:** Trained on diverse manuscripts, notably the Syro-Coptic Synaxarion of Aleppo and various historical colophons.
- **RESULT:** The first public model for Garshuni-Arabic achieved an exceptional ***2.97% CER*** (Model ID 578529) across 326 training pages and 79,679 words. Developed in collaboration with Coptic monks from the Monasteries of Saint Mina and Saint Anthony the Great, Egypt.



IV. Vienna 'Porta Orientis' Pilot & Global Capacity Building

THE EAST SYRIAC MODEL:
Expanding the project's script coverage, the "Porta Orientis" pilot aims to build the first public East Syriac HTR model on Transkribus.

GLOBAL PARTNERSHIP:
This initiative establishes an exclusive data-sharing and capacity-building partnership with the Syriac manuscripts libraries in the Middle East.

CAPACITY BUILDING:
By training a core cohort of heritage graduate students culminating in an intensive "Scribe Challenge" hackathon in Vienna, this pilot generates the high-fidelity Ground Truth necessary to process historically important Syriac texts.



V. Conclusion & Open Access Resources

The utility of HTR for complex Semitic scripts has successfully transitioned from theoretical potential to a highly accurate, practical reality. By democratizing these tools, this project demonstrates that community-driven, editor-verified datasets are essential to guarantee the digital future of global philology.

Acknowledgements

Supported by the FWF funded projects "Identifying Scattered Puzzles of Syriac Liturgy" (DOI: 10.55776/P36649) and the Cluster of Excellence EurAsian Transformations (Grant-DOI: 10.55776/COE8). Deep gratitude is extended to the Transkribus development team, the Monasteries of St. Mina and St. Anthony, the Syriac Groups in the IMAFO Winter Schools of HTR, and all the collaborators who contributed to building and supporting these models.



Explore the Open Access Tools:

Vienna Gospels Searchable Website:
app.transkribus.org/en/sites/Syriac-Vienna-Gospels/

Serto Model (Vienna Syriac Gospels)

Estrangelo-Melkite Model (Jerusalem MS 36)

Garshuni-Arabic Model (Aleppo Synaxarion)

FWF Austrian Science Fund
EurAsia
IMAFO
ÖAW Österreichische Akademie der Wissenschaften
excellent = austria
Institut für Mittelalterforschung Abteilung Byzanz Forschung