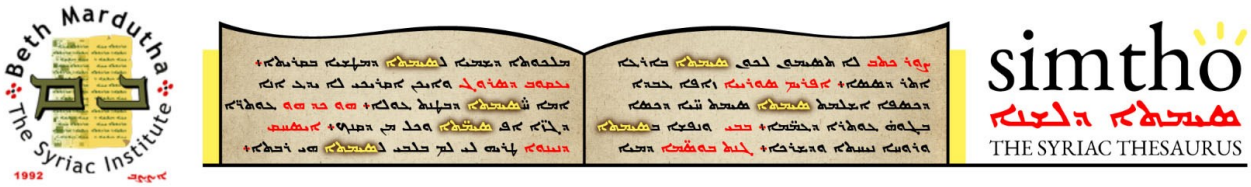


# Building an Electronic Syriac Corpus using OCR: Preserving and Digitizing Cultural Heritage—Launch of Simtho III

November 10, 2023

Registration required: https://bit.ly/Simtho3

**Sponsored by:** Institute for Advanced Study, Princeton and Beth Mardutho: The Syriac Institute.
**Conveners:** Sabine Schmidtke, George A. Kiraz and María Mercedes Tuya

The digitization of cultural heritage materials plays a crucial role in preserving and making accessible historical and linguistic resources. The Simtho corpus is a result of constructing an electronic Syriac corpus through the application of Optical Character Recognition (OCR) technology and correcting the OCR results in collaboration with young women and men in the Middle East who make up Beth Mardutho's MelthoLab team. Syriac, an ancient language with a rich literary and religious tradition, presents unique challenges due to its complex script. The proposed approach involves a multi-step process: digitization of physical books and manuscripts (by third parties), training custom OCR models for Syriac script recognition, and creating an annotated corpus for linguistic research. This project presents the culmination of this work, a 16 million-word corpus of Syriac texts. The resulting electronic corpus provides a digital repository of Syriac texts, enabling scholars, linguists, and historians to access, study, and analyze these valuable resources. Additionally, the project contributes to advancing OCR technology by developing and fine-tuning models for non-Latin scripts, potentially benefiting other languages with complex scripts and limited digital presence. The combination of Syriac studies and OCR technology showcases the potential of interdisciplinary collaboration in preserving and exploring cultural heritage in the digital age.

**Program (**Time Zone: Eastern Time - US & Canada **)**

| | |
|---|---|
| 12:00 pm | Welcome, Sabine Schmidtke and Maria Mercedes Tuya |
| 12:10 pm | Simtho overview, George A. Kiraz |
| 12:20 pm | From Simtho II to Simtho III, Michael Ethington (Università degli Studi di Padova) |
| 12:35 pm | Preparing Texts for Simtho, Eva Rodrigo (Hebrew University, Jerusalem) |
| 12:50 pm | 10-minute break |
| 1:00 pm | **Guest Speakers:** |
| | A Case Study on Timothy I, Michael Penn (Stanford University) |
| | Counting on Simtho: Some Challenges and Benefits for non-Specialists, James C. Wolfe (University of Michigan) |
| | **What's Next:** |
| 1:30 pm | Tokenization and NLP, Slavomír Čéplö (CERES, Ruhr-Universität Bochum) |
| 1:45 pm | Feedback and Ways Forward, Daniel Stoekl (École Pratique des Hautes Études) |
| 2:00 pm | Q & A |
| 2:30 pm | Closing remarks |