AI4Culture: Towards Multilingual Access for Cultural Heritage Data

Tom Vanallemeersch, Sara Szoc, Laurens Meeus
CrossLang NV, Franklin Rooseveltlaan 348/8, 9000 Gent, Belgium
{firstname.lastname}@crosslang.com

Abstract

The AI4Culture project (2023–2025), funded by the European Commission, and involving a 12-partner consortium led by the National Technical University of Athens, develops a platform serving as an online capacity building hub for AI technologies in the cultural heritage (CH) sector, enabling multilingual access to CH data. It offers access to AI-related resources, including openly labelled datasets for model training and testing, deployable and reusable tools, and capacity building materials. The tools are aimed at optical character recognition (OCR) for printed and handwritten documents, subtitle generation and validation, machine translation (MT), and metadata enrichment via image information extraction and semantic linking. The project also customises these tools to enhance interface and component usability. We illustrate this with technology that corrects OCR output using language models and adapts it for MT.

1 Introduction

The AI4Culture project aims to develop an online capacity building hub for AI technologies in the cultural heritage (CH) sector. This innovative platform seeks to make CH data more accessible and understandable in today's multilingual digital era, by facilitating data sharing, promoting cultural content reuse and linking the vast European data space (which ensures data availability for economic, societal and research use) with CH institutions.

The project is funded by the European Commission (EC) and runs from April 2023 to March 2025. The consortium is led by the AILS Laboratory of the National Technical University of Athens (NTUA). The other consortium partners are the NTUA spin-off Datoptron and organisations specialised in digital CH (i.e. Europeana Foundation, European Fashion Heritage Association, the DigitGLAM unit at University of Leuven, and the company Datable), natural language processing (the companies CrossLang and Pangeanic, the MT Research Unit at Fondazione Bruno Kessler (FBK), the Digital Safety and Security Center of the Austrian Institute of Technology), online translation services (the company Translated), and media culture (Institute for Sound and Vision).

The AI4Culture platform, expected to launch its first version mid-2024, will offer access to AI-related resources, i.e. to openly labelled datasets for training and testing models, to deployable and reusable tools, and to capacity building materials on the use of these tools and datasets for training, testing, and evaluation. The platform targets CH students and professionals, data providers, researchers and AI model developers, amongst others. Towards the end of 2024, several workshops will be organised on the technologies involved.

2 Technology and language coverage

The tools made accessible through the platform relate to four technologies:

1. Multilingual text recognition in scanned printed and handwritten documents through optical character recognition (OCR), machine

translation (MT), and semi-automatic validation of transcriptions.

2. Automated generation of multilingual subtitles and semi-automatic subtitle validation.
3. Enrichment of CH metadata through information extraction from images (color detection, object detection) and semantic linking (e.g. named entities).
4. Generation of multilingual versions of CH metadata using MT.

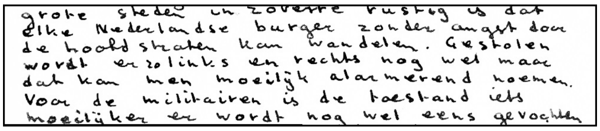
Accessibility of the above technologies will be achieved through online interfaces, application programming interfaces (APIs), which allow to send requests to various services, and docker images, which can be locally deployed. The interfaces include Transcribathon[1] (supporting the transcription of documents and the translation of transcriptions), Subbit![2] (supporting the editing of automatically generated subtitles), the interface of SAGE[3] (serving semantic annotation and generation of enrichments), and PECAT[4] (supporting validation and post-editing of automatic translations).

The software built during the project to achieve accessibility will be provided as open source. Moreover, this software focuses on the reuse of existing open-source tools, such as PERO-OCR.[5]

On the multilingual level, the transcription and subtitling services, as well the CH metadata, cover numerous languages, with a focus on EU official languages. The translation functionality uses various systems, including the Europeana Translate[6] and the EC's eTranslation engines.

3 Customisation of technologies

The tools made accessible on the platform are customised to increase the usability of interfaces and components. For instance, FBK's subtitling components incorporate the Whisper[7] pre-trained model for improving speech recognition. Another example is the post-correction component developed by Crosslang, which aims at enhancing both OCR and MT outputs of auto-generated transcriptions. This component particularly targets documents for which no (closely) matching specialised transcription engine exists in terms of language, time period, script of writing, etc.



grote steden in zoverre rustig is dat
elke Nederlandse burger zonder angst door
de hoofdstraten kan wandelen. Gestolen
wordt er zo links en rechts nog wel maar
dat kan men moeilijk alarmerend noemen.
Voor de militairen is de toestand iets
moeilijker en wordt nog wel eens gevochten

Figure 1: Part of Dutch letter from World War II

OCR	de hoofdstraten kan wandelen. Gestolen wordt ***errolinks*** en rechts nog
MT	can walk the main streets. ***Stolen it will be erroleft*** and right
Segmented	Gestolen wordt ***errolinks*** en rechts nog
MT	***Stolen*** is still happening on the left and right
Corrected	Gestolen wordt **er links** en rechts nog
MT	There is still **theft** left and right

Table 1: Effect of segmentation and word correction on OCR and MT output

The OCR post-correction component consist of several steps. First, it removes word-splitting hyphens at line breaks using a language model-based technique. Next, it segments the transcribed lines into sentences using an advanced rule-based approach. Finally, it performs word post-correction through either a basic method using lexicon and language model lookup, or through a computationally more demanding strategy that prompts a chatbot based on a large language model (LLM) to make corrections to the transcriptions.

Figure 1 shows a part of a Dutch letter from World War II.[8] Table 1 compares different OCR[9] and MT outputs[10] for a sentence from this letter, highlighting the effects of the post-correction component on OCR and MT results. The segmentation step leads to an enhancement of the MT output, consisting of a better handling of the non-existent Dutch word *errolinks* (ground truth *er zo links*). The word post-correction step (using the LLM-based method mentioned above) brings the Dutch word closer to its ground truth by replacing it with *er links*, thus further improving the MT output of the sentence (containing *theft* instead of *stolen*).

Acknowledgements: AI4Culture is funded by EC's DIGITAL programme (project 101100683).

[1] transcribathon.eu

[2] subbit.eu

[3] pro.europeana.eu/page/sage

[4] pangeanic.com/datasets-for-ai/ai-data-annotation-platform

[5] github.com/DCGM/pero-ocr

[6] pro.europeana.eu/project/europeana-translate

[7] github.com/openai/whisper

[8] zenodo.org/records/8108347 (we converted the background color to white for better contrast in the figure)

[9] Produced by the Text Titan I model from Transkribus (readcoop.eu/transkribus).

[10] Produced by Google Translate.