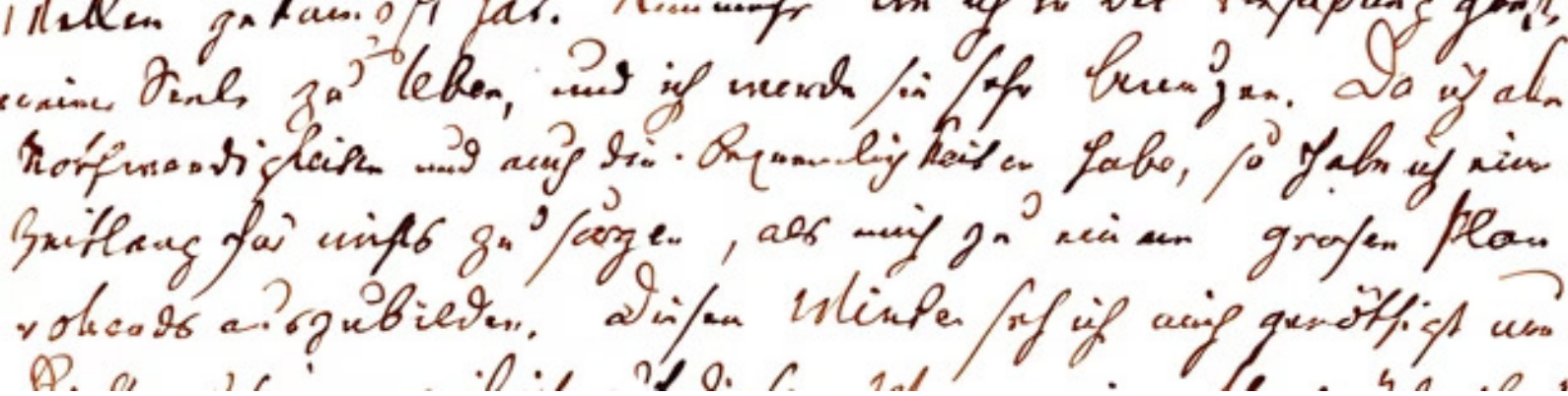


INTERCOML | INTERNATIONAL CONFERENCE ON DIGITAL HUMANITIES

Transcription of Czech administrative records written in Kurrent script into Latin script using machine learning, OCR, and AI

Author: Tomáš Martínek, Czech University of Life Sciences Prague, martinekt@pef.czu.cz

What Is Kurrent?

Kurrent is an old form of German cursive handwriting, widely used for centuries in administrative records and personal correspondence. While its origins trace back to the 16th century, its use for Czech language texts peaked around 1850 and continued until 1941 for German texts. The distinct letter forms and ligatures present a significant challenge for modern transcription efforts.

The Challenge of Czech Kurrent

Mixed Scripts

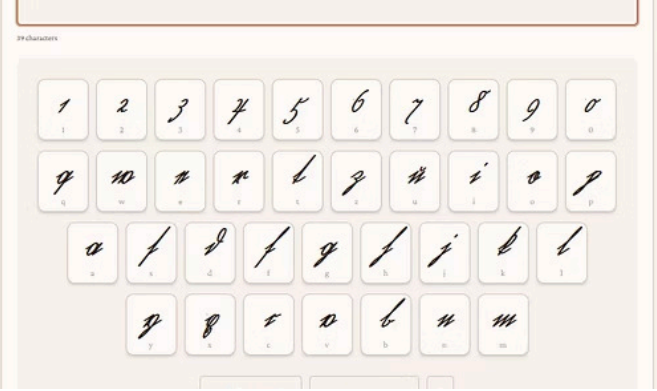
Many historical documents feature a blend of Kurrent and Humanistic scripts, often within the same sentence, making automated identification and transcription complex.

Individual Variation

The highly individualistic nature of Kurrent handwriting means that no two scribes wrote exactly alike, adding to the difficulty of creating a universal transcription model.

Training Data

A significant hurdle is the scarcity of high-quality, accurately transcribed training data required to effectively teach machine learning models to recognize and translate Kurrent script.



Importance and Existing Tools

Why This Matters Today

The transcription of historical Kurrent script holds significant value in the modern digital age, impacting accessibility, research, and data utility.

Digitization Wave

As historical archives and administrative records are increasingly digitized, efficient and accurate transcription methods are crucial for converting vast amounts of Kurrent script into searchable and accessible Latin script. This effort ensures the preservation and broader availability of invaluable historical data.

Business Value

Unlocking the content of these historical documents can provide profound insights across various domains, including agronomy, historical research, genealogy, and cultural studies. This accessibility can drive new research avenues, educational initiatives, and even commercial applications through the data. Current contracts may refer to historical land registers.

Metadata

Accurately transcribed texts allow for robust metadata extraction and enrichment. This process significantly improves the searchability, cross-referencing, and overall utility of historical documents, making it easier for researchers and the public to navigate and utilize complex historical datasets.

Existing Tools & Approaches

Several tools and methodologies currently address the challenges of historical script transcription, each with its strengths and specific applications in the field of digital humanities.

Transkribus

A platform for the automated recognition, transcription, and searching of historical documents, widely used by scholars for various historical scripts.

PERO OCR

An open-source OCR system developed by Brno University of Technology, known for its flexibility and ability to process diverse historical documents and layouts.

inkCapture

A specialized tool often used for handwriting recognition, offering capabilities to train models on specific hands or script styles for improved accuracy.

TrOCR-Kurrent

A model fine-tuned for Kurrent script transcription, leveraging transformer-based OCR for enhanced performance on this challenging historical script.

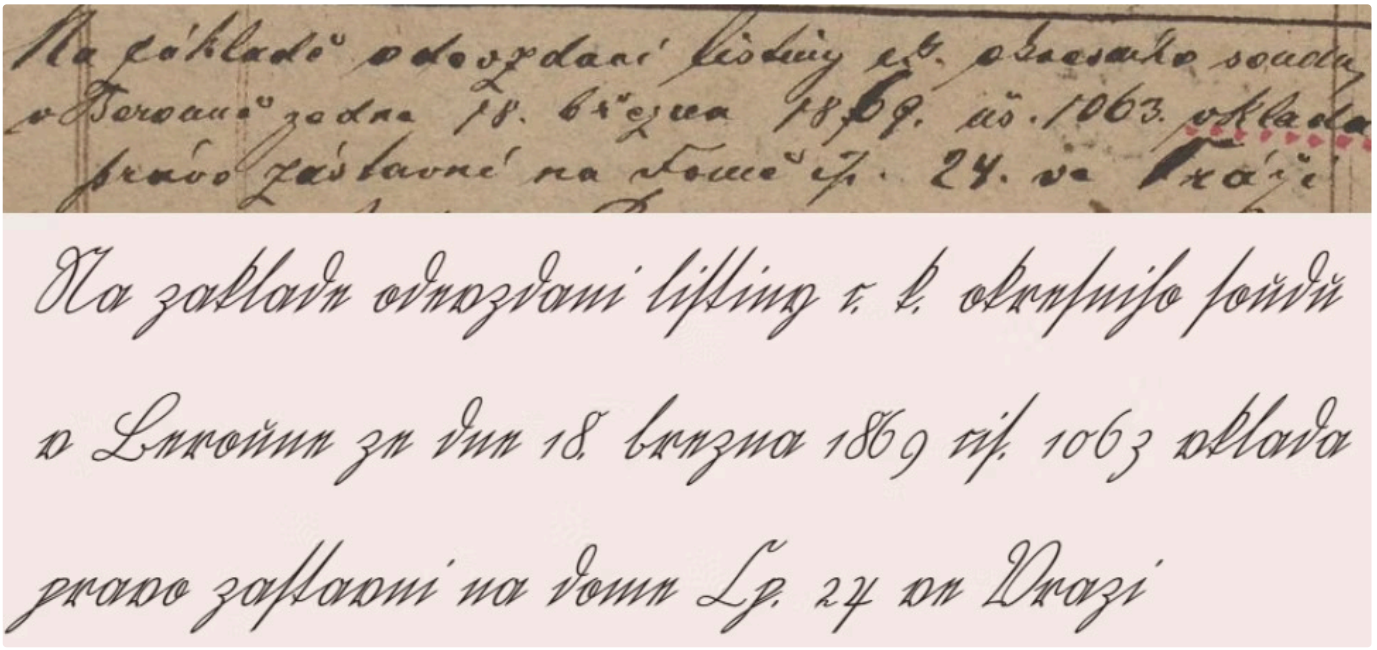


Figure: Example of Kurrent script from a historical administrative record (Source: Vráž).

Transcription of an entry from the Book of Assets and Liabilities of the village of Vráž into a standardized font, which after transcription reads "Based on the submission of document no. 1063 of the Imperial-Royal District Court in Beroun dated March 18, 1869, a lien is placed on house no. 24 in Vráž."

Online Tools and AI Models for Kurrent Transcription

The Online Transcription Tool

The online tool, **kurent.pridat.eu**, serves as a dedicated platform for transcribing Kurrent script. This tool supports various fonts, enabling users to convert historical administrative records and other documents into Latin script with accuracy and efficiency.

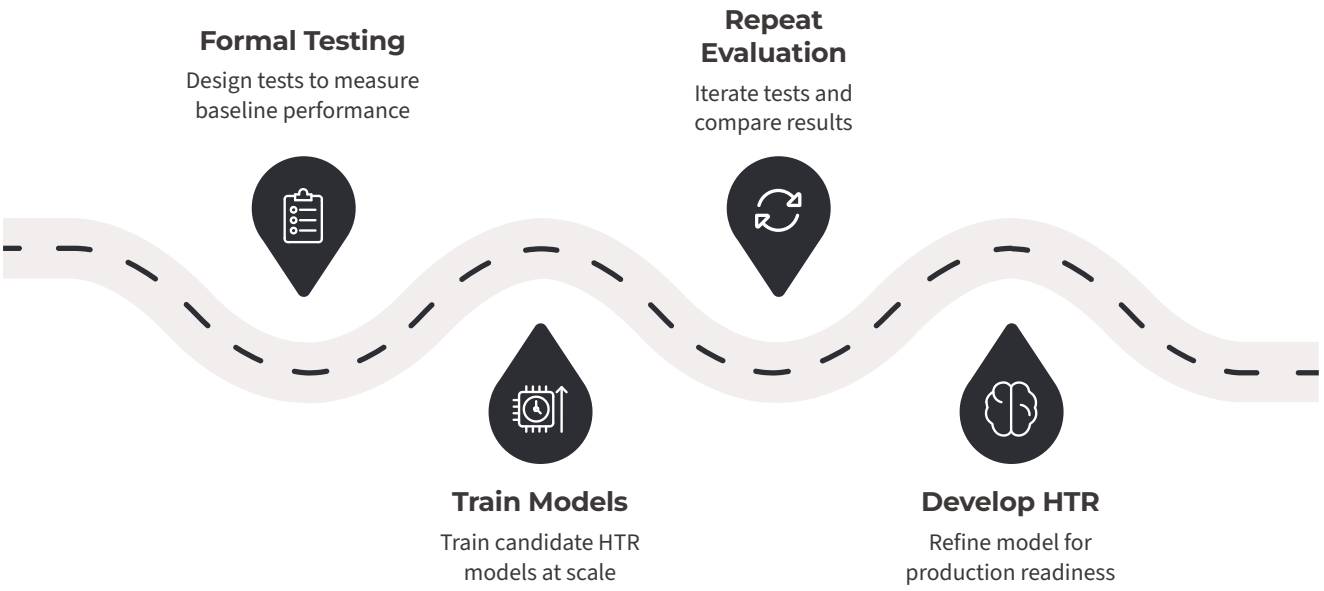
AI Models and other tools

For comparison and further analysis, several advanced AI models were utilized in the transcription efforts:

ChatGPT-5.4	**Gemini 3.1 Pro**	**Claude Sonnet 4.6**	**Kimi K2.5**
Mistral	**Nemotron 3 Super**	**xAI: Grok**	**Google Lens**

Research Roadmap and Vision

Research Roadmap



This roadmap outlines the systematic approach to developing and refining the transcription model, ensuring robustness and accuracy through iterative testing and training cycles.

Vision & Impact

New HTR Model

Development of a highly accurate Handwritten Text Recognition (HTR) model specifically tailored for Czech Kurrent script, capable of handling its unique complexities and variations.

At-Scale Transcription

Enable the transcription of vast quantities of historical administrative records and other documents written in Kurrent script, making previously inaccessible information digitally available and searchable.

Practical Applications

Facilitate new research opportunities in history, linguistics, and social sciences by providing rich, machine-readable datasets. Support genealogical studies and cultural heritage preservation efforts.

Ing. Tomáš Martínek, Ph.D., martinekt@pef.czu.cz

InterCoML 2026 · Prague, Czech Republic · April 27–30, 2026