The Patrologia Graeca Corpus

Chahan Vidal-Gorène, Bastien Kindt

The Patrologia Graeca Corpus

OCR, Annotation, and Open Release of Noisy Nineteenth-Century Polytonic Greek Editions

Chahan Vidal-Gorène[1,2], Bastien Kindt[3]
[1]École nationale des chartes-PSL (France) [2]Calfa (France)
[3]UCLouvain - CIOL Institut orientaliste (Belgium)
Contact: chahan.vidal-gorene@chartes.psl.eu

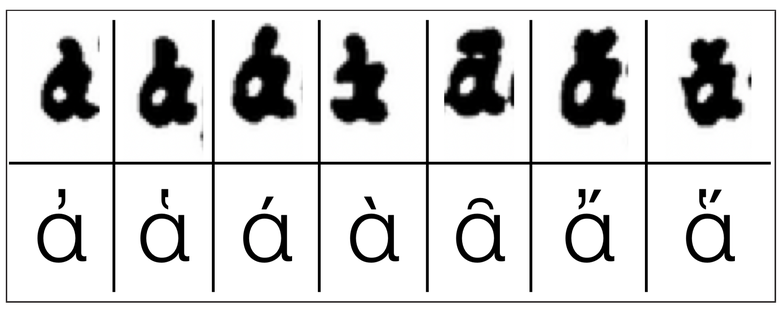
Context

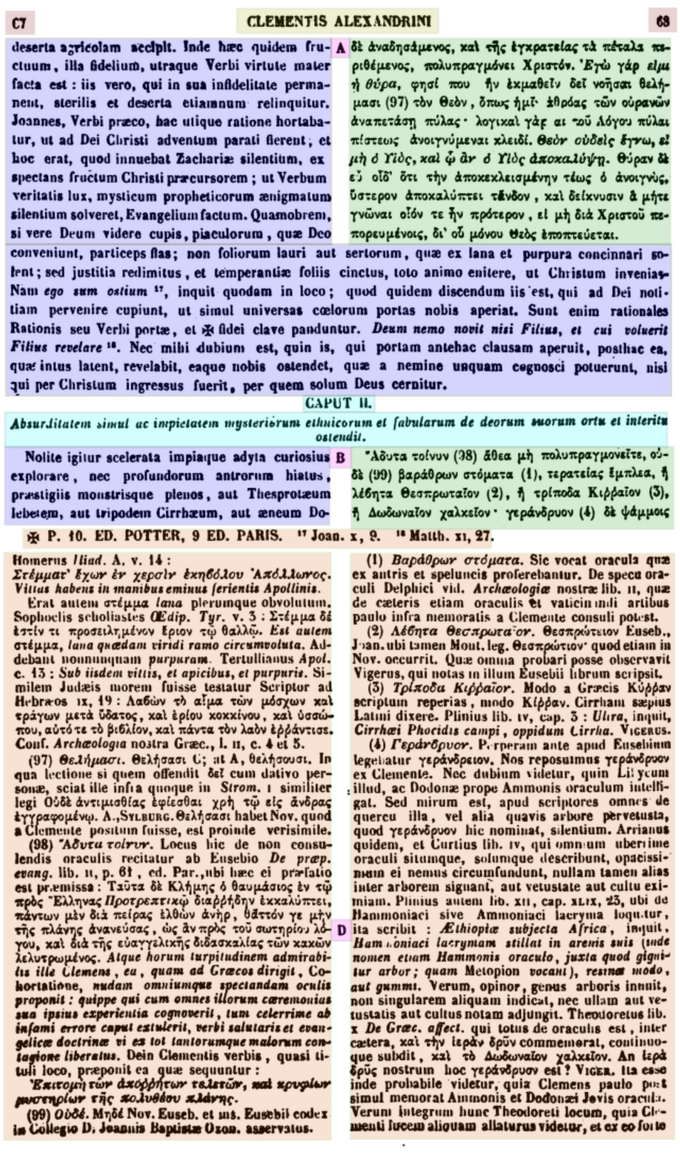
Objective: OCR, lemmatization and POS-tagging of the **still-unavailable 33 volumes** of the *Patrologia Graeca* (**post-classical and Byzantine texts**).

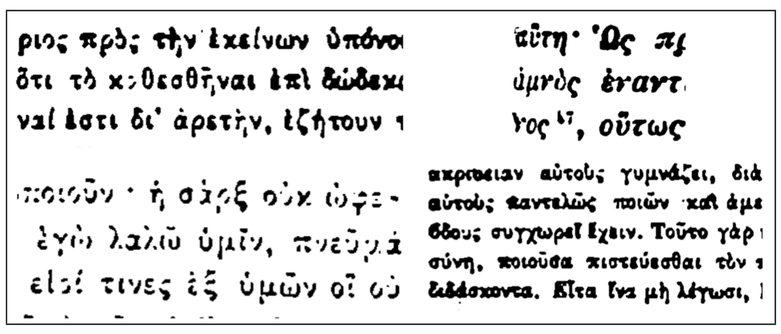
Challenge: Processing noisy 19th-century editions to **fill gaps left by scholarly databases** like the *Thesaurus Linguae Graecae (TLG)*.

Outcome: The first large-scale, open-access OCR and silver linguistic resource for Ancient Greek with **6M words** ($< 1.05\%$ of CER).

Challenges







Release

Searchable corpus:
`https://www.gregoriproject.com`

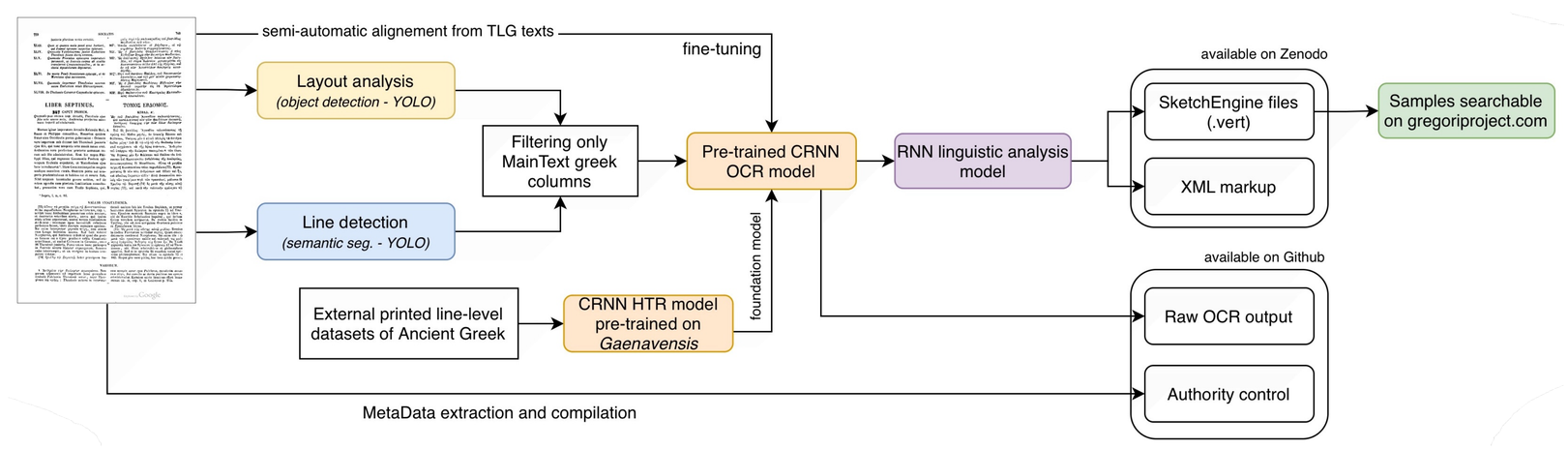
Full raw data (OCR and tagged texts, `.txt`/`.vert`):
`https://zenodo.org/records/15780625`

OCR ground truth:
`https://github.com/calfa-co/Patrologia-Graeca`

Text extraction and text analysis pipeline

To overcome the complex bilingual layouts (with crossing-lines) and degraded 19th-century typography, the pipeline relies on a **YOLO12s-detect model** for semantic zone detection and a **YOLO12s-seg model** for line segmentation. The text is then extracted using a **CRNN-based text recognition** (Vidal-Gorène *et al.*, 2021) pretrained with raw alignments from the TLG, then fine-tuned for the *Patrologia Graeca*, achieving a 1.05% Character Error Rate (CER).

Following text extraction, the pipeline applies a hybrid strategy combining the PIE architecture (Manjavacas *et al.*, 2019), already pretrained with GRE*g*ORI, and a **dictionary-based post-correction** (Kindt *et al.*, 2022).



Results and corpus coverage

Our fine-tuned OCR yields **1.05% CER** and **4.69% WER**, thanks to the **pretraining on real data** (≈11k lines). Layout and line detection achieve high precision (mAP50 ≥ 0.97). Notably, 80% of remaining errors are minor polytonic diacritic confusions rather than base character errors (no confusion observed at the lemmatization step). The resulting corpus comprises **6M lemmatized tokens** with full morphosyntactic annotation. t-SNE highlights the PG cluster (yellow) as a distinct semantic space driven by its specific vocabulary, a key for training new diachronically robust Ancient Greek language models.

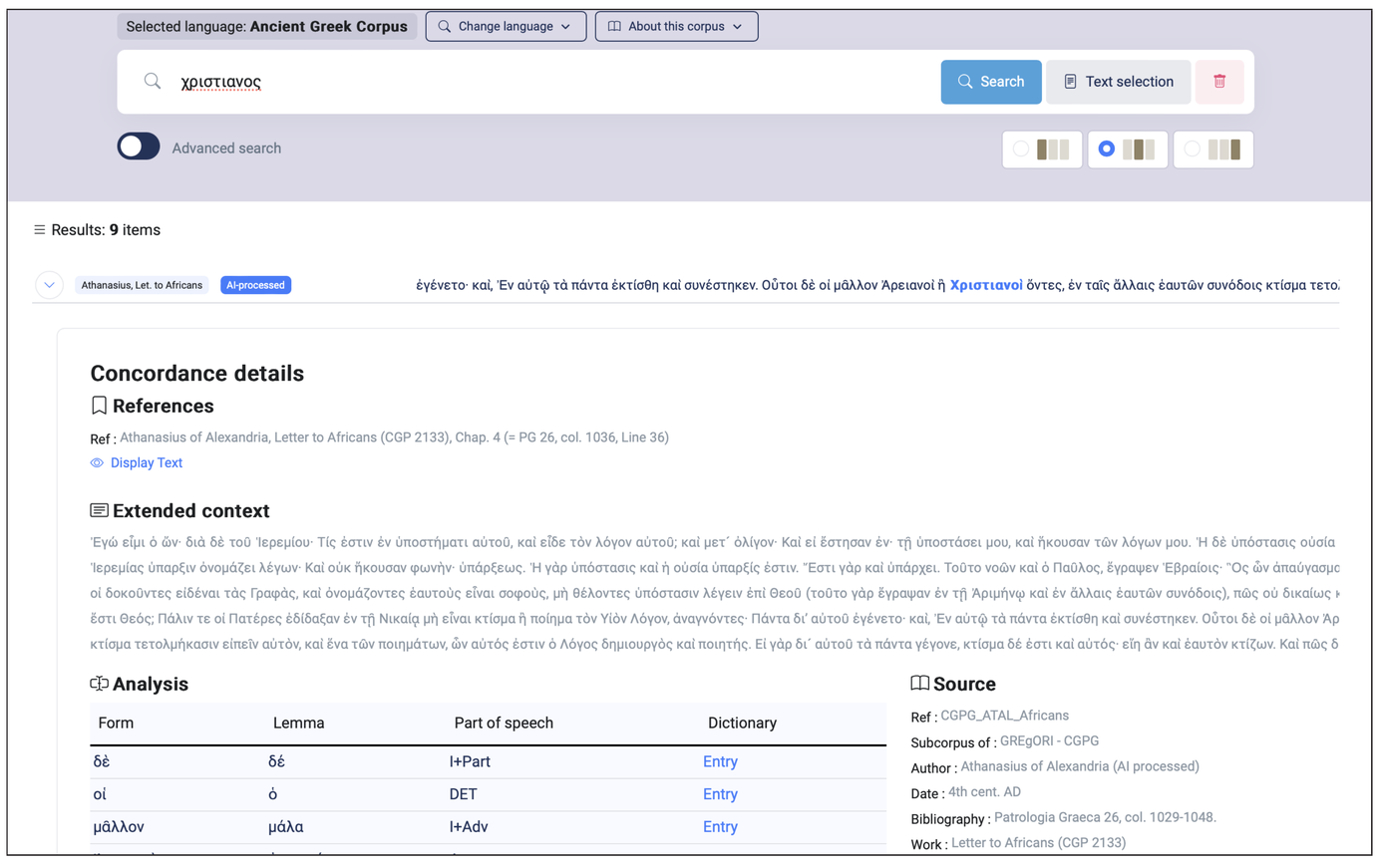
Text recognition

Model	CER	WER
Tesseract	11.57%	39.65%
Transkribus	6.14%	14.82%
Ours (PG)	**1.05%**	**4.69%**

Layout detection

Class	P	R	mAP50
Greek column	0.963	0.994	0.973
Latin column	0.969	1.000	0.982
Title	0.462	0.950	0.462
Marginalia	0.422	0.891	0.427
Line detection			
Detection	0.983	0.994	0.973
Reading order	0.980	—	—

t-SNE: PG forms distinct semantic cluster, reflecting GT rich vocabulary compared to existing datasets.



Lemma search in the corpus, displaying concordances, analyzes, context, metadata and linked to lexical resources

[1] Kindt *et al.*, *Analyse automatique du grec ancien par réseau de neurones [...]*, BABELAO, Louvain (2022)
[2] Vidal-Gorène *et al.*, *A modular and automated annotation platform [...]*, ICDAR, Geneva (2021)
[3] Manjavacas *et al.*, *Improving Lemmatization of Non-Standard Languages [...]*, AACL, Minneapolis (2019)

This research was supported by ASBL Byzantion, Calfa (Paris), UCLouvain (CIOL, FSS – Fondation Sedes Sapientiae, GREgORI Project, INCAL, RSCS), and private funders.
We thank Prof. Jean-Marie Auwers, Prof. Sébastien Moureau, and Dr. Véronique Somers (UCLouvain) for their contributions to data curation, annotation, and evaluation.