Optical Character Recognition of 19th Century Classical Commentaries: the Current State of Affairs

Matteo Romanello (UNIL)
Sven Najem-Meyer (EPFL)
Bruce Robertson (Mount Allison Univ.)

HIP’21 @ ICDAR – September 6, 2021, Lausanne (CH)

Summary

1. Introduction
2. Datasets
3. Evaluation
4. Discussion

Photo credits: © Fondation Hardt, Vandoeuvres, 2019

Introduction

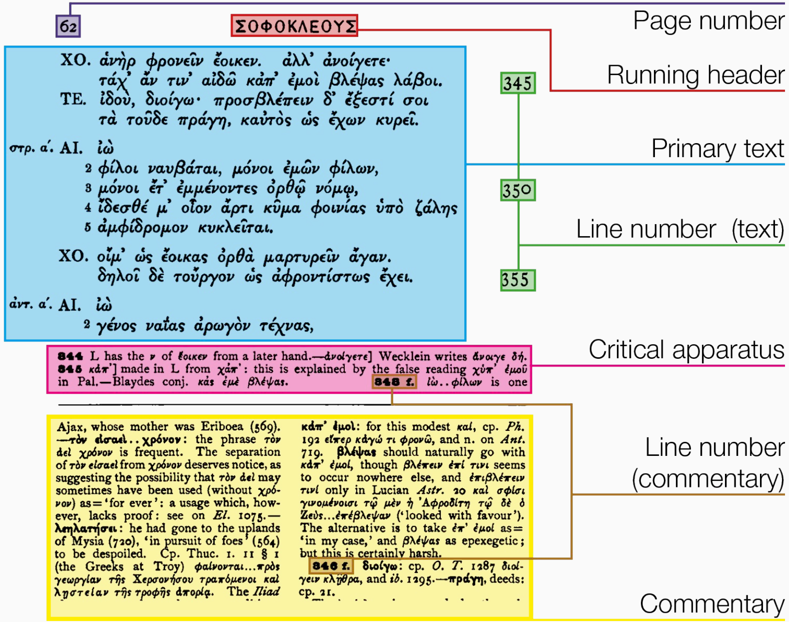
Classical Commentaries

Main forms of Classical scholarship:

- Editions
- Translations
- **Commentaries**

Century-long tradition of writing commentaries

Aims of a commentary: translate, make a text more accessible, contextualize, comment on history of text transmission, etc.



62 ΣΟΦΟΚΛΕΟΥΣ

ΧΟ. ἀνὴρ φρονεῖν ἔοικεν. ἀλλ' ἀνοίγετε·
τάχ' ἄν τιν' αἰδῶ κἀπ' ἐμοὶ βλέψας λάβοι.
ΤΕ. ἰδού, διοίγω· προσβλέπειν δ' ἔξεστί σοι
τὰ τοῦδε πράγη, καὐτὸς ὡς ἔχων κυρεῖ.

στρ. α'. ΑΙ. ἰὼ
2 φίλοι ναυβάται, μόνοι ἐμῶν φίλων,
3 μόνοι ἔτ' ἐμμένοντες ὀρθῷ νόμῳ,
4 ἴδεσθέ μ' οἷον ἄρτι κῦμα φοινίας ὑπὸ ζάλης
5 ἀμφίδρομον κυκλεῖται.

ΧΟ. οἴμ' ὡς ἔοικας ὀρθὰ μαρτυρεῖν ἄγαν.
δηλοῖ δὲ τοὔργον ὡς ἀφροντίστως ἔχει.

ἀντ. α'. ΑΙ. ἰὼ
2 γένος ναΐας ἀρωγὸν τέχνας,

844 L has the ν of ἔοικεν from a later hand.—ἀνοίγετε] Wecklein writes ἄνοιγε δή. **845** κἀπ'] made in L from χἀπ': this is explained by the false reading χὐπ' ἐμοῦ in Pal.—Blaydes conj. κἀς ἐμὲ βλέψας. **848 f.** ἰὼ..φίλων is one

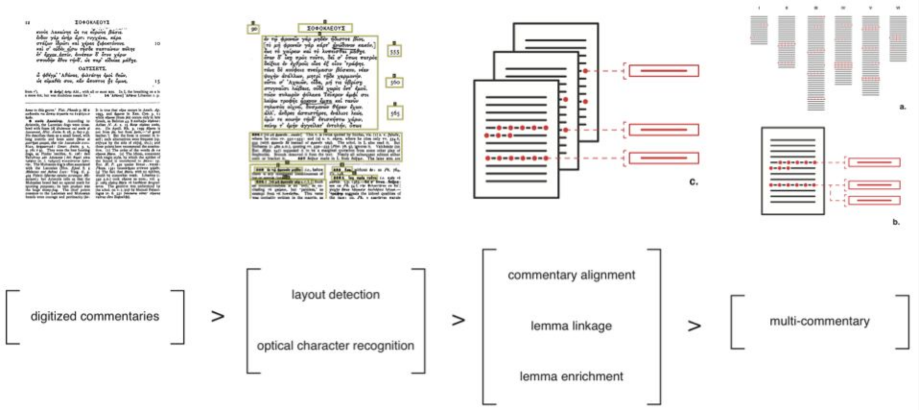
Ajax, whose mother was Eriboea (569). —**τὸν εἰσαεὶ..χρόνον**: the phrase τὸν ἀεὶ χρόνον is frequent. The separation of τὸν εἰσαεὶ from χρόνον deserves notice, as suggesting the possibility that τὸν ἀεὶ may sometimes have been used (without χρόνον) as='for ever': a usage which, however, lacks proof: see on *El.* 1075.—**λεηλατήσει**: he had gone to the uplands of Mysia (720), 'in pursuit of foes' (564) to be despoiled. Cp. Thuc. I. 11 § 1 (the Greeks at Troy) φαίνονται...πρὸς γεωργίαν τῆς Χερσονήσου τραπόμενοι καὶ λῃστείαν τῆς τροφῆς ἀπορίᾳ. The *Iliad*

κἀπ' ἐμοί: for this modest καί, cp. *Ph.* 192 εἴπερ κἀγώ τι φρονῶ, and n. on *Ant.* 719. **βλέψας** should naturally go with κἀπ' ἐμοί, though βλέπειν ἐπί τινι seems to occur nowhere else, and ἐπιβλέπειν τινὶ only in Lucian *Astr.* 20 καὶ σφίσι γινομένοισι τῷ μὲν ἡ 'Αφροδίτη τῷ δὲ ὁ Ζεὺς...ἐπέβλεψαν ('looked with favour'). The alternative is to take ἐπ' ἐμοὶ as= 'in my case,' and βλέψας as epexegetic; but this is certainly harsh.
846 f. **διοίγω**: cp. *O. T.* 1287 διοίγειν κλῇθρα, and *ib.* 1295.—**πράγη**, deeds: cp. 21.

The *Ajax* Multi-Commentary

Project goal: an epistemological study of *Ajax*’s commentaries.

A *digital multi-commentary* will allow to **read**, **compare** and **analyze** the entire commentary tradition of this tragedy.



[digitized commentaries] > [layout detection, optical character recognition] > [commentary alignment, lemma linkage, lemma enrichment] > [multi-commentary]

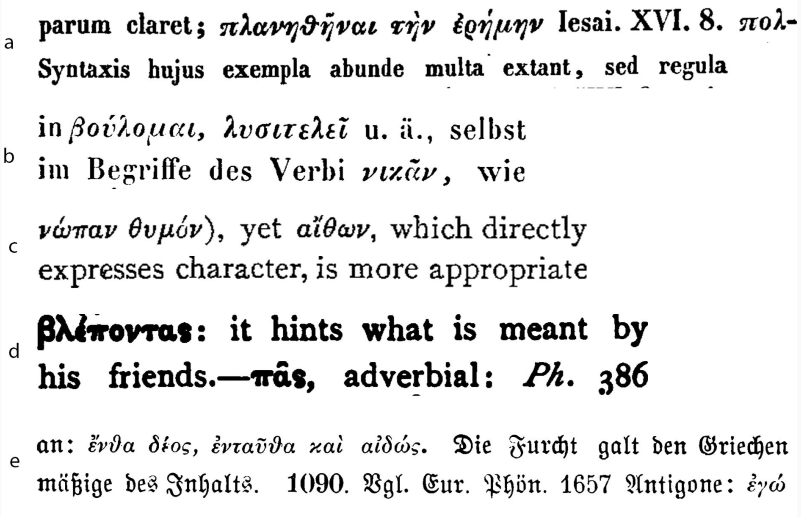
FNSNF
Swiss National Science Foundation

UNIL | Université de Lausanne

EPFL

Challenges for OCR

- Quality of digitized images
- Quantity of available training GT data
- Complexity of layouts
- Mix of Latin and polytonic Greek scripts
- Variety of typefaces for Greek



a
parum claret; *πλανηθῆναι τὴν ἐρήμην* Iesai. XVI. 8. *πολ-*
Syntaxis hujus exempla abunde multa extant, sed regula

b
in *βούλομαι, λυσιτελεῖ* u. ä., selbst
im Begriffe des Verbi *νικᾶν*, wie

c
νώπαν θυμόν), yet *αἴθων*, which directly
expresses character, is more appropriate

d
βλέποντας: it hints what is meant by
his friends.—πᾶς, adverbial: ***Ph.*** **386**

e
an: *ἔνθα δέος, ἐνταῦθα καὶ αἰδώς.* Die Furcht galt den Griechen
mäßige des Inhalts. 1090. Vgl. Eur. Phön. 1657 Antigone: *ἐγώ*

Datasets

GT4HistComment

Ground truth data for OCR of historical classical commentaries

Five 19th century commentaries on Sophocles' *Ajax*

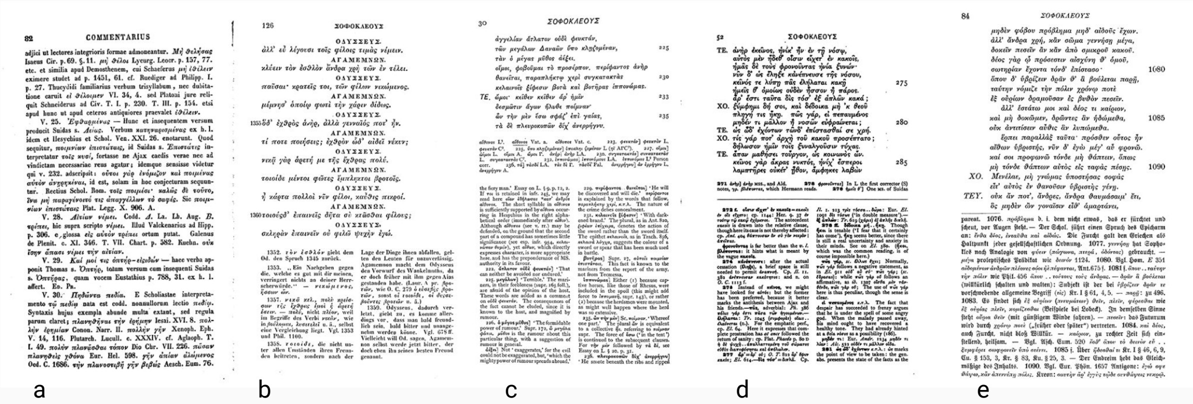
Languages: German, English, Latin

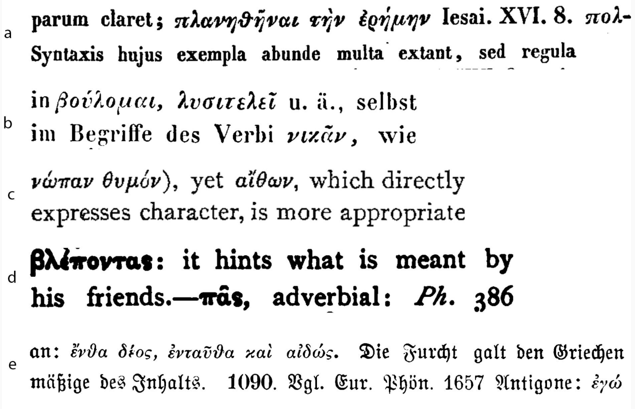
Total of 3,356 lines

GT used for evaluation & retraining

CC By License

https://github.com/AjaxMultiCommentary/GT-commentaries-OCR





parum claret; *πλανηθῆναι τὴν ἐρήμην* Iesai. XVI. 8. *πολ-*
Syntaxis hujus exempla abunde multa extant, sed regula

in *βούλομαι, λυσιτελεῖ* u. ä., selbst
im Begriffe des Verbi *νικᾶν*, wie

νώπαν θυμόν), yet *αἴθων*, which directly
expresses character, is more appropriate

βλέποντας: it hints what is meant by
his friends.—πᾶς, adverbial: ***Ph.*** **386**

an: *ἔνθα δέος, ἐνταῦθα καὶ αἰδώς.* Die Furcht galt den Griechen
mäßige des Inhalts. 1090. Vgl. Eur. Phön. 1657 Antigone: *ἐγώ*

a. Lobeck (1835)

b. Schneidewin (1853)

c. Campbell, (1881)

d. Jebb (1896)

e. Wecklein (1894)

PoGreTra

Polytonic Greek Training Data from Historic Texts

OCR GT data + pre-trained Kraken classifiers

Supported typefaces: Porson and “German-serifs”

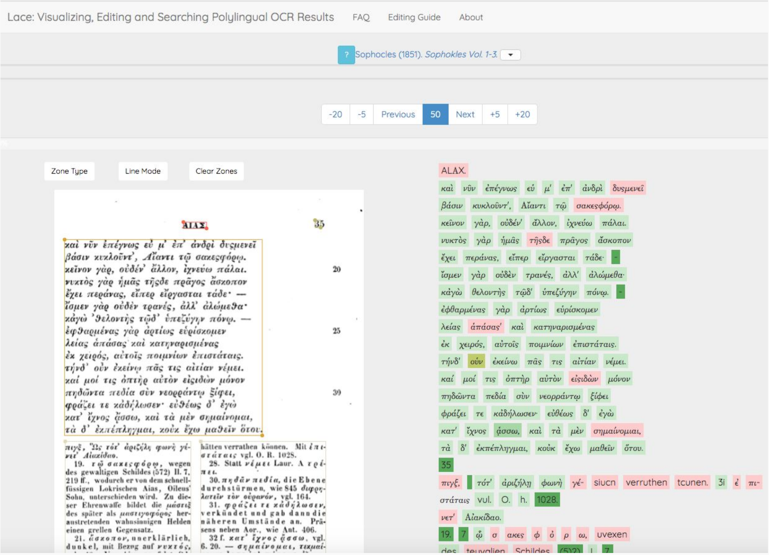
Total of 31,972 lines (6,607 Porson + 25,365 German-serifs), and ~300k tokens

https://doi.org/10.5281/zenodo.4774200

Open Greek & Latin + First Thousand Years of Greek

Ongoing effort to create an open corpus with at least one edition of every Greek work composed between Homer and 250 CE

To date, over 22M words of manually transcribed Classics primary sources were released


Lace: Visualizing, Editing and Searching Polylingual OCR Results
FAQ
Editing Guide
About
Sophocles (1851). Sophokles Vol. 1-3.
-20
-5
Previous
50
Next
+5
+20
Zone Type
Line Mode
Clear Zones

ALAX.
καὶ νῦν ἐπέγνως εὔ μ' ἐπ' ἀνδρὶ δυςμενεῖ
βάσιν κυκλοῦντ', Αἴαντι τῷ σακεςφόρῳ.
κεῖνον γὰρ, οὐδέν' ἄλλον, ἰχνεύω πάλαι.
νυκτὸς γὰρ ἡμᾶς τῆςδε πρᾶγος ἄσκοπον
ἔχει περάνας, εἴπερ εἴργασται τάδε· -
ἴσμεν γὰρ οὐδὲν τρανές, ἀλλ' ἀλώμεθα·
κἀγὼ θελοντὴς τῷδ' ὑπεζύγην πόνῳ. -
ἐφθαρμένας γὰρ ἀρτίως εὑρίσκομεν
λείας ἁπάσας' καὶ κατηναρισμένας
ἐκ χειρός, αὐτοῖς ποιμνίων ἐπιστάταις.
τήνδ' ούν ἐκείνω πᾶς τις αἰτίαν νέμει.
καί μοί τις ὀπτὴρ αὐτὸν εἰςιδὼν μόνον
πηδῶντα πεδία σὺν νεορράντῳ ξίφει
φράζει τε κἀδήλωσεν· εὐθέως δ' ἐγὼ
κατ' ἴχνος ᾄσσω, καὶ τὰ μὲν σημαίνομαι,
τὰ δ' ἐκπέπληγμαι, κοὐκ ἔχω μαθεῖν ὅτου.
35
πιγξ, τότ' ἀριζήλη φωνὴ γέ- siucn verruthen tcunen. 3i ἐ πι-
στάταις vul. O. h. 1028.
νετ' Αἰακίδαο.
19. 7 ῷ σ ακες φ ὀ ρ ω, uvexen
des teuvalien Schildes (5)2) | 7

https://github.com/brobertson/Lace2

Evaluation

Pipeline 1 : Tesseract/OCR-D

OCR-D. Complete framework for:

- Pre-processing images
- OLR
- OCR
- Export to various formats
- Post-processings

+

Tesseract. Pre-trained models for:

- English, German, French...
- Fraktur
- Latin
- Polytonic Greek
- GT4HistOCR

Multi-models confidence-based voting available

Pipeline 2: Kraken+Ciaconna

Ciaconna:

- **Training**. Relies on **Kraken** to train models on custom data.
- **Data**. Data acquired in the context of Open Greek and Latin (OGL) (PoGreTra)
- **Post-processing.**
  - De-hyphenation
  - Diacritics correction
  - Spell Checking

Evaluation settings

Metrics:

- Normalized Levenshtein distance (NLD) = character accuracy = 1 - Character Error Rate (CER)
- F1-score : bag of words for TP, FP, TN and FN.

Unicode:

- Combined diacritic-main form (“NFC”) : ᾶ instead of ἁ → 0% NLD
- Decomposed form (“NFD”) : ῀ α instead of ῾ α → 50% NLD

Evaluation tool

- PRImA TextEval-like (Bag of word-based)
- OCLR/evaluation (coordinate-based)

Experiment 1: Base vs re-trained Kraken+Ciaconna.

Table. Base versus re-trained models' results by commentary.

Commentary Additional data (chars)	**Lobeck** +16084	**Schneidewin** +16113	**Jebb** +19141
Metric	NLD	NLD	NLD
Kraken+Ciaconna (base)	0.89	0.83	0.88
Kraken+Ciaconna (retrained)	**0.91**	**0.91**	**0.91**

General results by commentary

Table. Character accuracy by model and by commentary.

Commentary	**Lobeck**	**Schneidewin**	**Campbell**	**Jebb**	**Wecklein**
Calamari GT4Hist	0.63	0.72	0.73	0.69	0.68
Tesseract	0.89	**0.92**	**0.95**	**0.92**	0.95
Kraken+Ciaconna (base)	0.89	0.83	0.93	0.88	**0.95**
Kraken+Ciaconna (retrained)	**0.91**	0.91	-	0.91	-

General results by region type

Table. Weight-averaged (±STD) character accuracy by model and by region type

Region	**Global**	**Greek**	**Commentary**	**Low-Greek**	**App. Crit.**	**Structured**	**Numbers**
Nb. of chars (% Greek)	51186 (29%)	6657 (92%)	23825 (23%)	13322 (2%)	2062 (43%)	3371 (34%)	693 (0%)
Calamari GT4Hist	.70±.04	.16±.05	.73±.04	.95±.04	.54±.12	.66±.01	.77±.26
Tesseract	**.93±.02**	.87±.05	**.92±.02**	**.99±.00**	.88±.01	**.93±.01**	**.87±.13**
Kraken+Ciaconna	.92±.02	**.93±.04**	.89±.05	.96±.01	**.93±.00**	.93±.02	.87±.17

Discussion

One pipeline to rule ‘em all?

Commentary sections with high density of polytonic Greek:

- Tesseract/OCR-D **87%** vs Kraken + Ciaconna **93%**

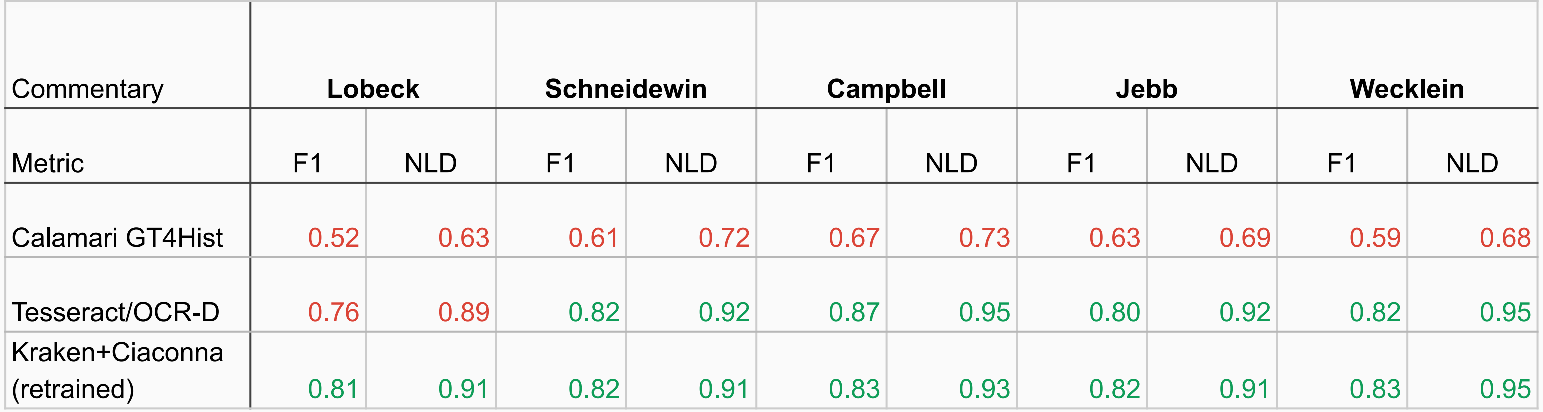
Commentary sections predominantly in Latin script:

- Tesseract/OCR-D **91.8%** vs Kraken + Ciaconna **91.6%**

Character accuracy on mixed script documents lower than SoTA on single-script docs:

- Tesseract/OCR-D 93%
- Kraken + Ciaconna 92%
- Polytonic Greek (Kiessling 2019) 99.2%
- Latin-script historical documents (Wick et al. 2018) 98-99%

Is the OCR fit for NLP?



Commentary	Lobeck		Schneidewin		Campbell		Jebb		Wecklein	
Metric	F1	NLD	F1	NLD	F1	NLD	F1	NLD	F1	NLD
Calamari GT4Hist	0.52	0.63	0.61	0.72	0.67	0.73	0.63	0.69	0.59	0.68
Tesseract/OCR-D	0.76	0.89	0.82	0.92	0.87	0.95	0.80	0.92	0.82	0.95
Kraken+Ciaconna (retrained)	0.81	0.91	0.82	0.91	0.83	0.93	0.82	0.91	0.83	0.95

- Topic modelling, vector space analysis, collocations, authorial attribution
  - OCRed texts with F-score >= 0.8 (Hill & Hengchen 2019)
- Sentence segmentation, named entity recognition, dependency parsing
  - OCRed texts with NLD > 0.9 (van Strien et al. 2020)

Thanks!

To contact us:

matteo.romanello@unil.ch

sven.najem-meyer@epfl.ch

broberts@mta.ca