OCR for non-Roman Scripts

Dale J. Correa
Adriana Cásarez

What is OCR?

Optical character recognition

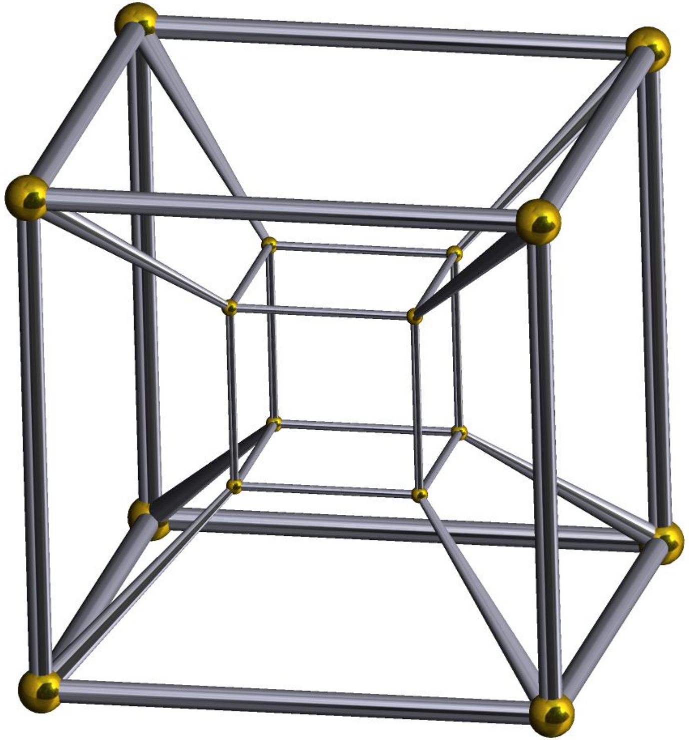
Method for making texts machine-readable so they can be used and manipulated in a variety of ways

Early OCR intended as accessible technology

Variations:

- Optical word recognition (words vs. characters)
- Intelligent character/word recognition (handwriting)

Tesseract



What is it really:

An OCR engine developed at HP between 1984 and 1994; unveiled to much fanfare in 1995

Tesseract: how does it work?

Page layout analysis: binary image with text regions

Connected component analysis: stores outlines of components

White on black (inverse) text

Blobs

Blobs become text lines, lines are analyzed for text by character or word

Two passes:

1) Recognize teach word, gather training data so it gets better as it goes down the page
2) Check top of page with new knowledge gained from pass 1

Final phase resolves fuzzy spaces and small text issues

Roman and non-Roman scripts

Roman script: graphic signs based on the letters of the classical Latin alphabet

Non-Roman script: anything that varies or differs from the Roman alphabet(s)

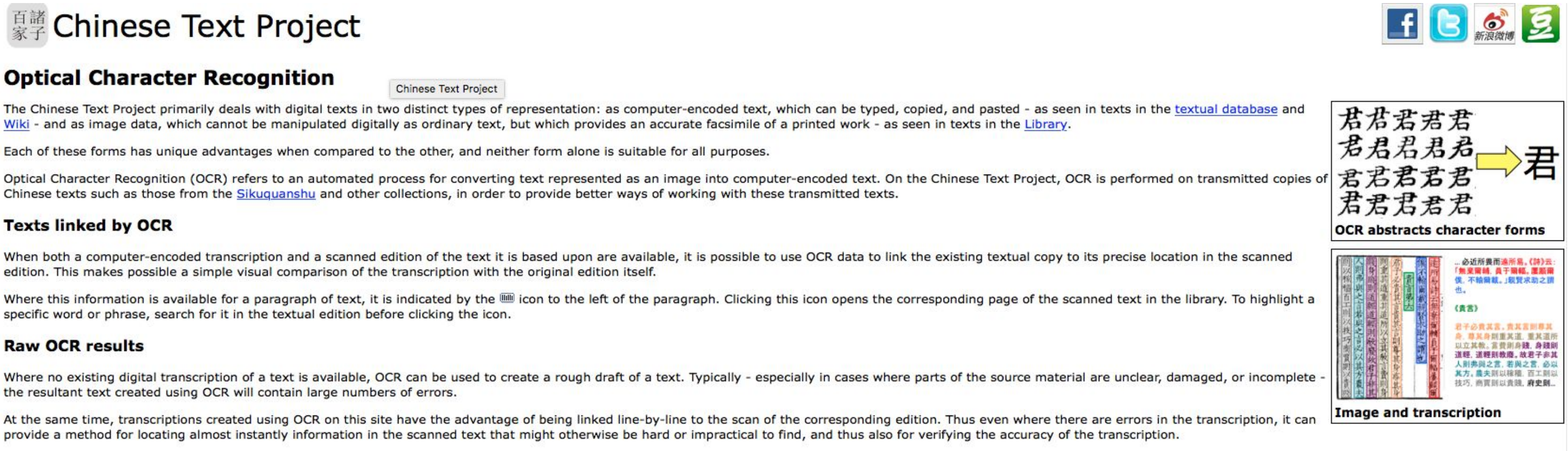
Directionality: some scripts flow LTR, others RTL; some flow horizontally, others vertically

*bidirectionality is a persistent issue in DS/DH work

Some scripts use spaces as word dividers, some do not

Some languages use multiple characters to form words, others use one character per word

Example: Chinese Text Project



百諸家子 Chinese Text Project

Optical Character Recognition

Chinese Text Project

The Chinese Text Project primarily deals with digital texts in two distinct types of representation: as computer-encoded text, which can be typed, copied, and pasted - as seen in texts in the textual database and Wiki - and as image data, which cannot be manipulated digitally as ordinary text, but which provides an accurate facsimile of a printed work - as seen in texts in the Library.

Each of these forms has unique advantages when compared to the other, and neither form alone is suitable for all purposes.

Optical Character Recognition (OCR) refers to an automated process for converting text represented as an image into computer-encoded text. On the Chinese Text Project, OCR is performed on transmitted copies of Chinese texts such as those from the Sikuquanshu and other collections, in order to provide better ways of working with these transmitted texts.

OCR abstracts character forms

Texts linked by OCR

When both a computer-encoded transcription and a scanned edition of the text it is based upon are available, it is possible to use OCR data to link the existing textual copy to its precise location in the scanned edition. This makes possible a simple visual comparison of the transcription with the original edition itself.

Where this information is available for a paragraph of text, it is indicated by the ▥ icon to the left of the paragraph. Clicking this icon opens the corresponding page of the scanned text in the library. To highlight a specific word or phrase, search for it in the textual edition before clicking the icon.

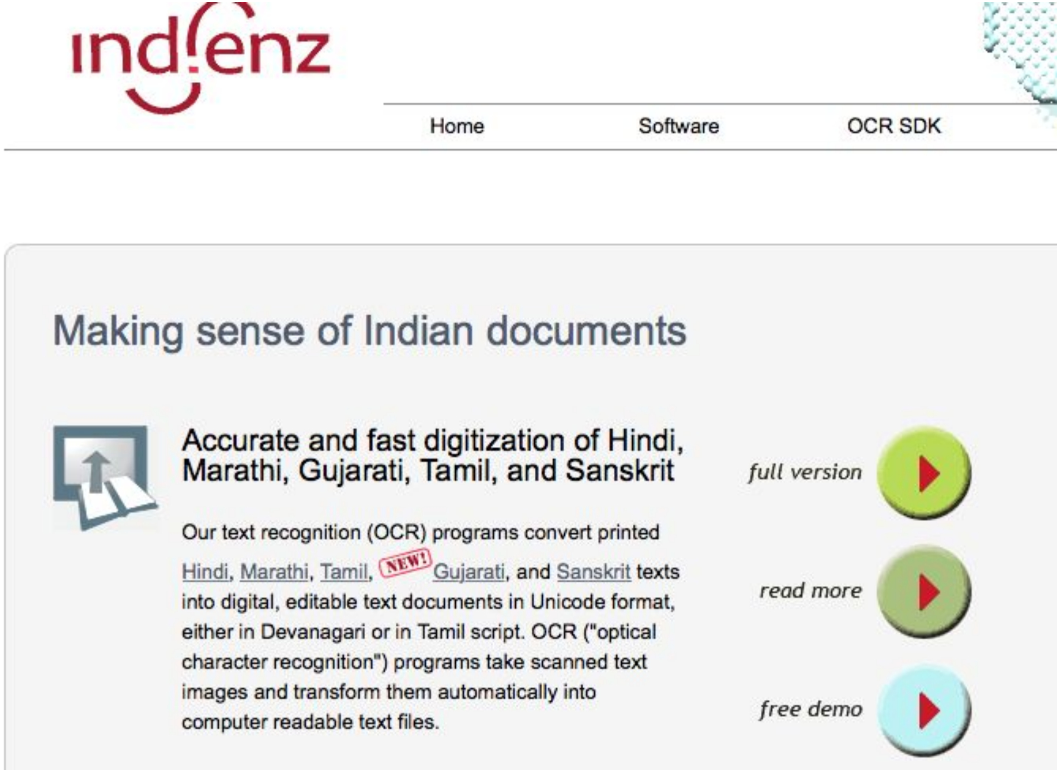
Image and transcription

Raw OCR results

Where no existing digital transcription of a text is available, OCR can be used to create a rough draft of a text. Typically - especially in cases where parts of the source material are unclear, damaged, or incomplete - the resultant text created using OCR will contain large numbers of errors.

At the same time, transcriptions created using OCR on this site have the advantage of being linked line-by-line to the scan of the corresponding edition. Thus even where there are errors in the transcription, it can provide a method for locating almost instantly information in the scanned text that might otherwise be hard or impractical to find, and thus also for verifying the accuracy of the transcription.

Example: indsenz

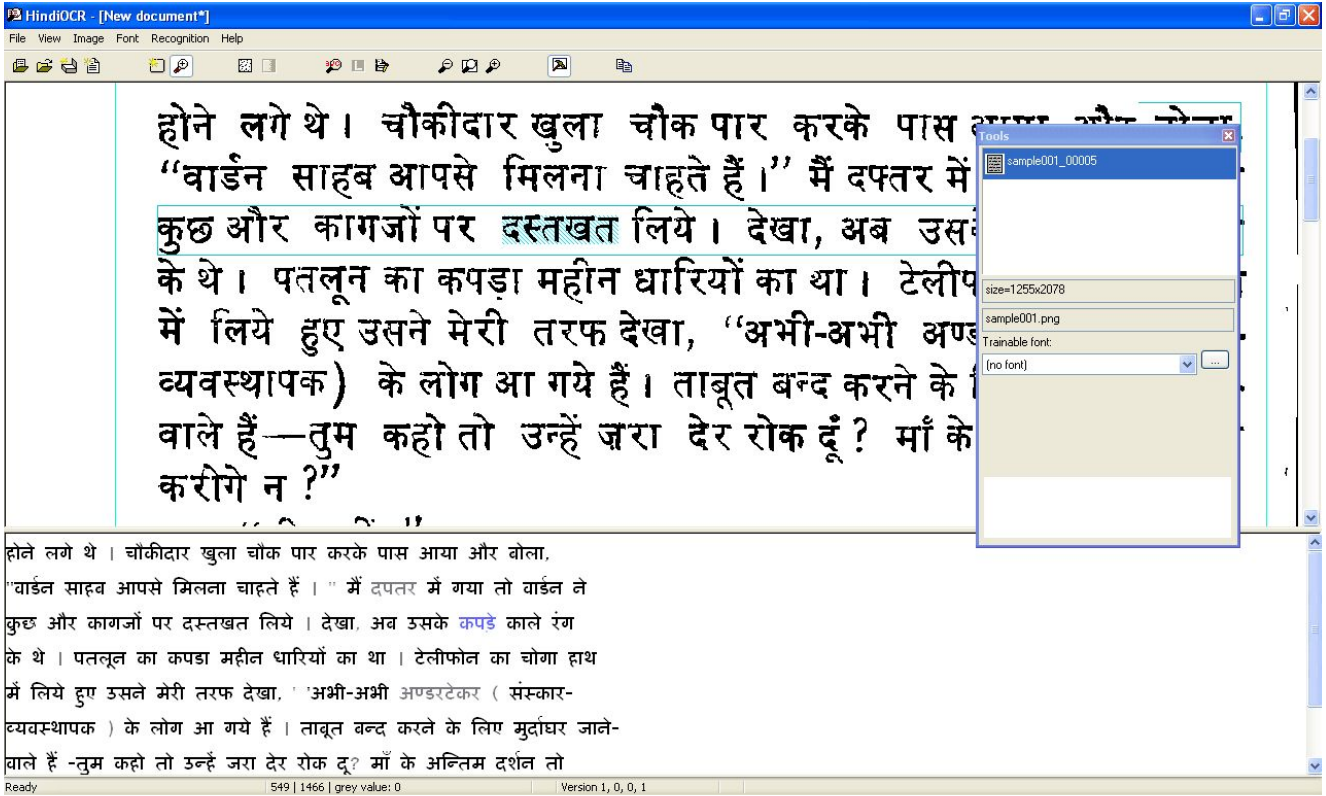


Making sense of Indian documents

Accurate and fast digitization of Hindi, Marathi, Gujarati, Tamil, and Sanskrit

Our text recognition (OCR) programs convert printed Hindi, Marathi, Tamil, NEW! Gujarati, and Sanskrit texts into digital, editable text documents in Unicode format, either in Devanagari or in Tamil script. OCR ("optical character recognition") programs take scanned text images and transform them automatically into computer readable text files.

Example: SanskritOCR



होने लगे थे। चौकीदार खुला चौक पार करके पास
“वार्डन साहब आपसे मिलना चाहते हैं।” मैं दफ्तर में
कुछ और कागजों पर दस्तखत लिये। देखा, अब उस
के थे। पतलून का कपड़ा महीन धारियों का था। टेलीप
में लिये हुए उसने मेरी तरफ देखा, “अभी-अभी अण्
व्यवस्थापक) के लोग आ गये हैं। ताबूत बन्द करने के
वाले हैं—तुम कहो तो उन्हें ज़रा देर रोक दूं? माँ के
करोगे न ?”

size=1255x2078

sample001.png

Trainable font:

(no font)

होने लगे थे । चौकीदार खुला चौक पार करके पास आया और बोला,
"वार्डन साहव आपसे मिलना चाहते हैं । " मैं दपतर में गया तो वार्डन ने
कुछ और कागजों पर दस्तखत लिये । देखा, अव उसके कपड़े काले रंग
के थे । पतलून का कपडा महीन धारियों का था । टेलीफोन का चोगा हाथ
में लिये हुए उसने मेरी तरफ देखा, ' 'अभी-अभी अण्डरटेकर (संस्कार-
व्यवस्थापक) के लोग आ गये हैं । तावूत वन्द करने के लिए मुर्दाघर जाने-
वाले हैं -तुम कहो तो उन्हें जरा देर रोक दू? माँ के अन्तिम दर्शन तो

549 | 1466 | grey value: 0

Example: kraken.re

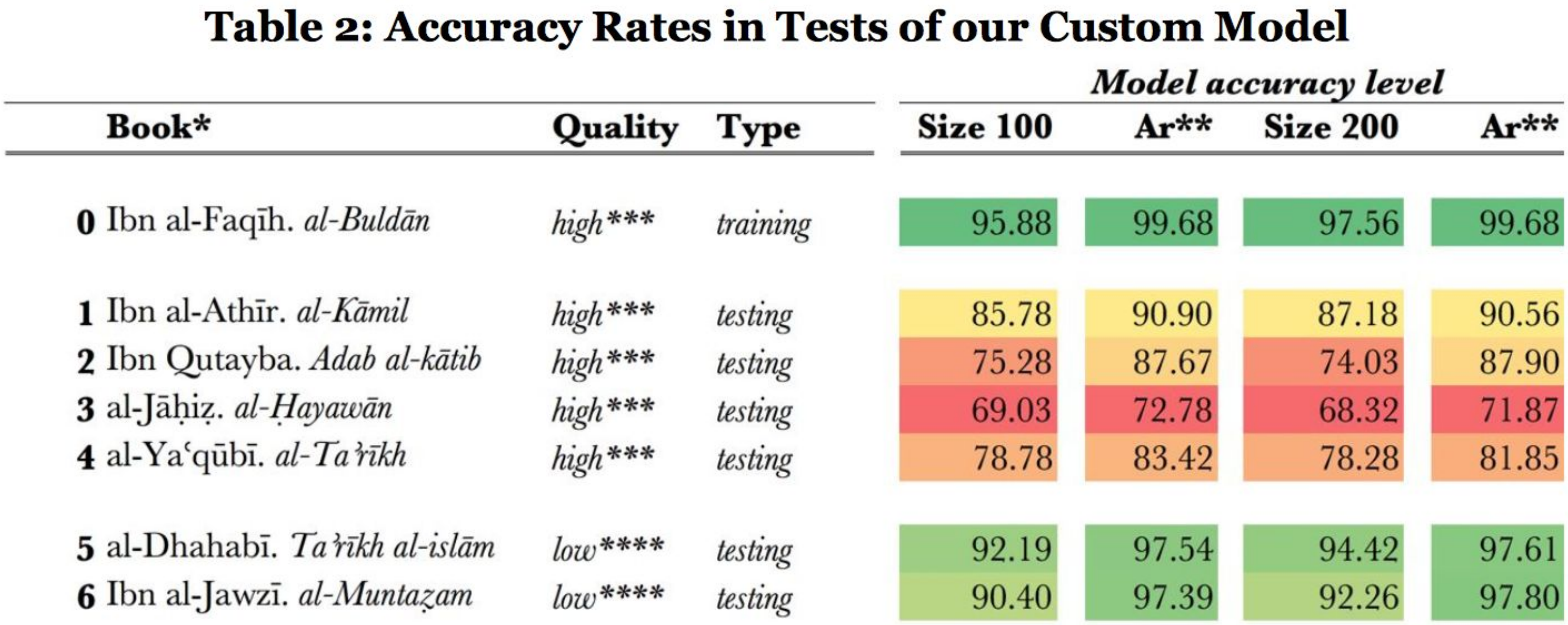


Table 2: Accuracy Rates in Tests of our Custom Model

			Model accuracy level			
Book*	**Quality**	**Type**	**Size 100**	**Ar****	**Size 200**	**Ar****
0 Ibn al-Faqīh. *al-Buldān*	*high****	*training*	95.88	99.68	97.56	99.68
1 Ibn al-Athīr. *al-Kāmil*	*high****	*testing*	85.78	90.90	87.18	90.56
2 Ibn Qutayba. *Adab al-kātib*	*high****	*testing*	75.28	87.67	74.03	87.90
3 al-Jāḥiẓ. *al-Ḥayawān*	*high****	*testing*	69.03	72.78	68.32	71.87
4 al-Yaʿqūbī. *al-Taʾrīkh*	*high****	*testing*	78.78	83.42	78.28	81.85
5 al-Dhahabī. *Taʾrīkh al-islām*	*low*****	*testing*	92.19	97.54	94.42	97.61
6 Ibn al-Jawzī. *al-Muntaẓam*	*low*****	*testing*	90.40	97.39	92.26	97.80

Projects Using Tesseract

Image Processing

Tesseract automatically processes the image but improving quality can yield better results

- Garbage In, Garbage Out

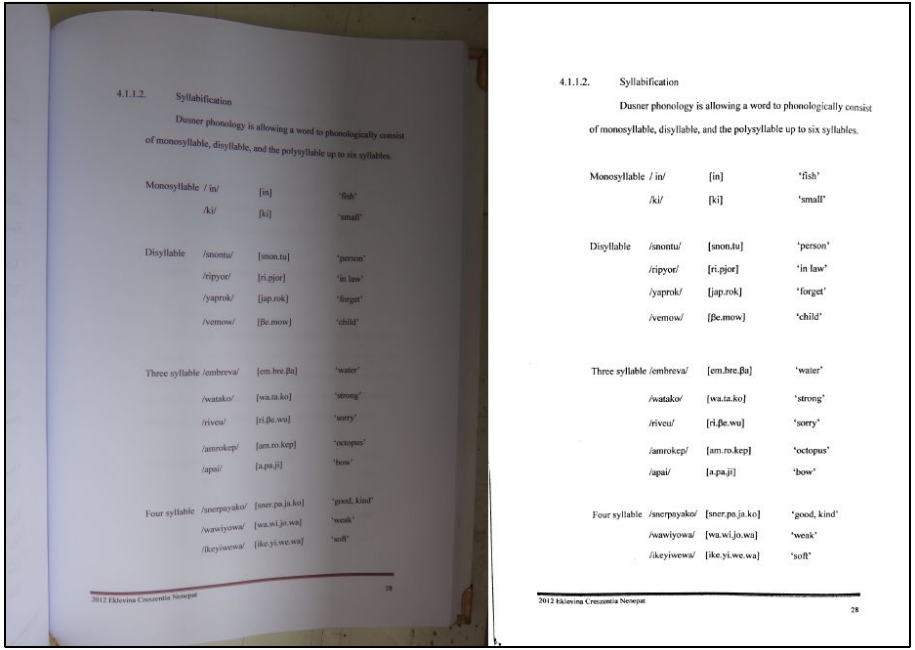
Image Editors

- Open Source (ImageJ, ImageMagick, OpenCV)

Tesseract’s ImproveQuality Page

Image Processing

Dewarping



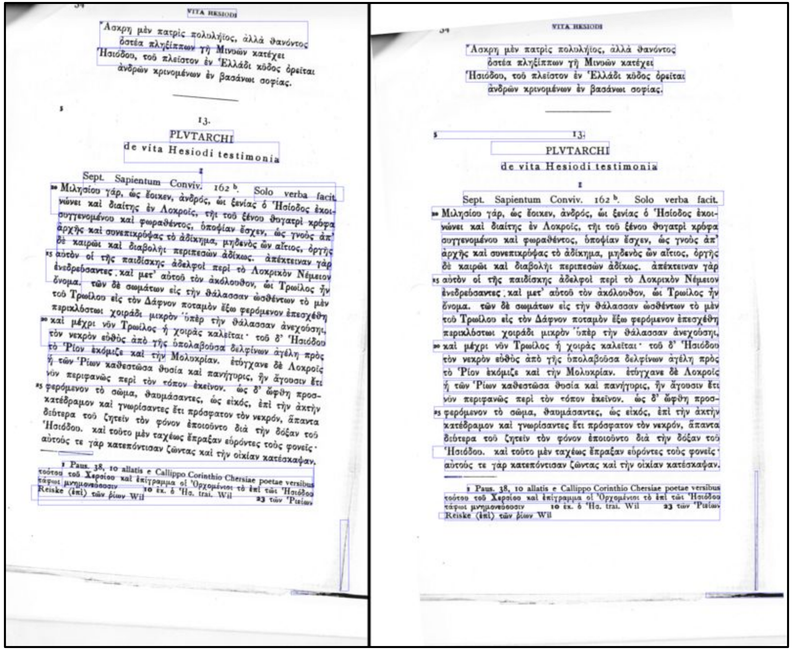
4.1.1.2. Syllabification

Dusner phonology is allowing a word to phonologically consist of monosyllable, disyllable, and the polysyllable up to six syllables.

Monosyllable	/ in/	[in]	'fish'
	/ki/	[ki]	'small'
Disyllable	/snontu/	[snon.tu]	'person'
	/ripyor/	[ri.pjor]	'in law'
	/yaprok/	[jap.rok]	'forget'
	/vemow/	[βe.mow]	'child'
Three syllable	/embreva/	[em.bre.βa]	'water'
	/watako/	[wa.ta.ko]	'strong'
	/riveu/	[ri.βe.wu]	'sorry'
	/amrokep/	[am.ro.kep]	'octopus'
	/apai/	[a.pa.ji]	'bow'
Four syllable	/snerpayako/	[sner.pa.ja.ko]	'good, kind'
	/wawiyowa/	[wa.wi.jo.wa]	'weak'
	/ikeyiwewa/	[ike.yi.we.wa]	'soft'

2012 Eklevina Creszentia Nenepat 28

Deskewing



34 VITA HESIODI

Ἄσκρη μὲν πατρὶς πολυλήιος, ἀλλὰ θανόντος
ὀστέα πληξίππων γῆ Μινυῶν κατέχει
Ἡσιόδου, τοῦ πλεῖστον ἐν Ἑλλάδι κῦδος ὀρεῖται
ἀνδρῶν κρινομένων ἐν βασάνωι σοφίας.

13.

PLVTARCHI

de vita Hesiodi testimonia

1

Sept. Sapientum Conviv. 162 b. Solo verba facit. Μιλησίου γάρ, ὡς ἔοικεν, ἀνδρός, ὧι ξενίας ὁ Ἡσίοδος ἐκοινώνει καὶ διαίτης ἐν Λοκροῖς, τῆι τοῦ ξένου θυγατρὶ κρύφα συγγενομένου καὶ φωραθέντος, ὑποψίαν ἔσχεν, ὡς γνοὺς ἀπ' ἀρχῆς καὶ συνεπικρύψας τὸ ἀδίκημα, μηδενὸς ὢν αἴτιος, ὀργῆς δὲ καιρῶι καὶ διαβολῆι περιπεσὼν ἀδίκως. ἀπέκτειναν γὰρ αὐτὸν οἱ τῆς παιδίσκης ἀδελφοὶ περὶ τὸ Λοκρικὸν Νέμειον ἐνεδρεύσαντες, καὶ μετ' αὐτοῦ τὸν ἀκόλουθον, ὧι Τρωίλος ἦν ὄνομα. τῶν δὲ σωμάτων εἰς τὴν θάλασσαν ὠσθέντων τὸ μὲν τοῦ Τρωίλου εἰς τὸν Δάφνον ποταμὸν ἔξω φερόμενον ἐπεσχέθη περικλύστωι χοιράδι μικρὸν ὑπὲρ τὴν θάλασσαν ἀνεχούσηι, καὶ μέχρι νῦν Τρωίλος ἡ χοιρὰς καλεῖται· τοῦ δ' Ἡσιόδου τὸν νεκρὸν εὐθὺς ἀπὸ γῆς ὑπολαβοῦσα δελφίνων ἀγέλη πρὸς τὸ Ῥίον ἐκόμιζε καὶ τὴν Μολυκρίαν. ἐτύγχανε δὲ Λοκροῖς ἡ τῶν Ῥίων καθεστῶσα θυσία καὶ πανήγυρις, ἣν ἄγουσιν ἔτι νῦν περιφανῶς περὶ τὸν τόπον ἐκεῖνον. ὡς δ' ὤφθη προσφερόμενον τὸ σῶμα, θαυμάσαντες, ὡς εἰκός, ἐπὶ τὴν ἀκτὴν κατέδραμον καὶ γνωρίσαντες ἔτι πρόσφατον τὸν νεκρόν, ἅπαντα δεύτερα τοῦ ζητεῖν τὸν φόνον ἐποιοῦντο διὰ τὴν δόξαν τοῦ Ἡσιόδου. καὶ τοῦτο μὲν ταχέως ἔπραξαν εὑρόντες τοὺς φονεῖς· αὐτούς τε γὰρ κατεπόντισαν ζῶντας καὶ τὴν οἰκίαν κατέσκαψαν.

1 Paus. 38, 10 allatis e Callippo Corinthio Chersiae poetae versibus τούτου τοῦ Χερσίου καὶ ἐπίγραμμα οἱ Ὀρχομένιοι τὸ ἐπὶ τῶι Ἡσιόδου τάφωι μνημονεύουσιν 10 ἐκ. ὁ Ἡσ. trai. Wil 23 τῶν Ῥείων Reiske (ἐπὶ) τῶν βίων Wil

Image Processing

Keep it Simple:

Microsoft **Photos**,
Mac OS’ **Preview**

Michael always says, K. I. S. S. Keep It Simple, Stupid. Great advice,
hurts my feelings every time.

Tesseract Installation

Tesseract Installation, but really:

Use Chrome

https://github.com/tesseract-ocr/tesseract/wiki

→ for Windows: https://github.com/UB-Mannheim/tesseract/wiki

Choose 64-bit

Open installed files

“Do you want to allow this app from an unknown publisher to make changes to your device?” YES

https://github.com/UB-Mannheim/tesseract/wiki

Tesseract Installation Continued

Select English > Next > I Agree > “Install for anyone using this computer,” Next

> Select Additional script data (download) and Additional language data (download), Next

>Change location from Program Files to Desktop, Next > Install

*Error with Laos language file, click OK

*Error with equ file, click OK

“Completed,” Next > Finish

Congratulations!

Tesseract: how to use it

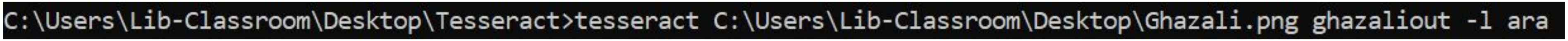
First, clean up desktop by

1) Creating a folder called Tesseract on desktop (right click; New; folder)
2) Drag all Tesseract files into this folder

Go to Start menu, find Tesseract in Program list

Click on Console under Tesseract list

Tesseract: your first command



```
C:\Users\Lib-Classroom\Desktop\Tesseract>tesseract C:\Users\Lib-Classroom\Desktop\Ghazali.png ghazaliout -l ara
```

Tesseract: your first command continued

```
tesseract [filename--drag into window] [brief filename]out -l [language/script code]
```

Ghazali (Arabic) example:

```
tesseract C:\Users\Lib-Classroom\Desktop\Ghazali.png ghazaliout -l ara
```

You will find “ghazaliout” as a file in your Tesseract folder, it is the OCR’d text of the ghazali image

Tesseract: try more languages and scripts

Chinese: -l chi_sim

Russian: -l rus

Devanagari: -l script\devanagari vs. -l hin+eng

Useful Commands

Basic Command Template:

```
tesseract imagename outputbase [-l lang]
```

Example:

```
tesseract myscan.png out -l chi_simp
```

Useful Commands Continued

-h, --help
Show help message.

--help-extra
Show extra help for advanced users.

--help-psm
Show page segmentation modes.

--help-oem
Show OCR Engine modes.

-v, --version
Returns the current version of the tesseract(1) executable.

--list-langs
List available languages for tesseract engine.
Can be used with --tessdata-dir *PATH*.

--print-parameters
Print tesseract parameters.