

Correzione collaborativa di risorse digitali acquisite tramite OCR

Federico Boschetti*

`federico.boschetti@yahoo.com`

* *CNR-ILC*

Pisa, 13 Giugno 2017

Introduzione

Il miglioramento delle prestazioni dell'OCR ha permesso di spostare l'attenzione dai soli testi digitali alle edizioni digitali. Infatti nella creazione delle prime collezioni digitali, prefazioni, introduzioni, indici, bibliografie, note, apparati critici e varianti testuali presenti in edizioni differenti abitualmente non venivano presi in considerazione.

Edizione critica

ΑΙΣΧΥΛΟΥ	
ἐχνοροῖι πεποιθώς στιφελαις ἐφέταις, χρυ- σογόνοι γενεᾶς ἰσθίους φῶς.	80
κυνέου δ' ὄμμασι λιύσσων φοίνου δέργμα δράκοντος, πολύχερ και πολυαιύτης, λύριον θ' ἄρμα δίκων, ἐπάγει δουρικλύτος ἀν- δράσι τοξόδαμον Άρη.	85 [στρ. β.]
δόκιμος δ' οὔτις ὑποστάς μεγάλω βρέματι φωτῶν ἐχνοροῖς ἔρκεσιν εἴργειν ἀμαχον κῆμα θαλάσσης· ἀπράσιτος γάρ ὁ Περσῶν στρατὸς ἀλκίφρων τε λαός.	90 [ἀντ. β.]
θεῖθεν γάρ κατὰ Μοῖρ' ἐκράτησεν τὸ παλαιόν, ἐπέκρηψε δὲ Πέρσας πολύμους πυργουδαίτους διέπειν ἔπιποχάρμας τε κλύους πάλαιον τ' ἀναστάσεις.	95 100 [στρ. γ.]
ἔμαλον δ' εὐρυπόροιο θαλάσσης πολαινομένης πνεύματι λάβρω	110 [ἀντ. γ.]

78 ἐχνοροῖι οὐ suprascripto M: ὀχροροῖι ut videtur Φ: cf. Sg: utrumque PQ στιφῆλοι codd. 79-80 χρυσογόνοι memorent ut v.l. (5 καὶ βλῆται) ΣΜΣΦQΡ: χρυσογόνοι codd. 80 ἰσθίον M 81 κύνεου] κυασίν Blomfield: cf. Theb. 78 82 φοίνου M: φοίνου fere tell. δῆρμα M 83 πολυαιύτης A 84 εὐκων M Trl.: δασίων vel ἀσίριον tell. 86 ἄρμ PVQ: cf. Theb. 45 89 ὀχροροῖς Φ: cf. 78 et Ag. 44 93-100 ponit vn. 101-114 tralicit O. Müller 99 πάλαιον S' fere codd.: corr. Byz. 101 θαλάσσης M Φ πολυαιουμένη M¹V

ΠΕΡΣΑΙ

ἔσορᾶν πόντιον ἄλοος, πίσινοι λεπτοδόμοι πεί- μμασι λα- σπόροις τε μηχαναῖς.	105 114
δολόμητην δ' ἀπάταν θεοῦ τίς ἀνήρ θνατός ἀλύξει; τίς ὁ κρασιγῶ ποδὶ πηδή- ματος εὐπέτεός ἀνάσσων; φιλόφρων γάρ (ποτι)σφαίνου- σα τὸ πρώτον παραίγει βροτῶν εἰς ἄρκυας Άτα, τόθεν οὐκ ἔστιν ὑπὲρ θνα- τῶν ἀλύξαντα φυγεῖν.	93 [μικροδ.] 95 110 100
ταυτὰ μοι μελαγχίτων φρήν ἀμύσσοται φόβω— ὁἶ Περαικοῦ στρατεύματος— τοῦδε μὴ πάλαι πύθη- ται, κένανθρον μὲν' ἄστν Σουσιδος·	[στρ. δ.] 116
καὶ τὸ Κισσιῶν πόλιον· ἀντίθουπον ἄσεται, ὁἶ, τοῦτ' ἔπος γυναικοπλη- θῆς ὁμιλος ἀπύων, βουσινοῖς δ' ἐν πέλοισι πίσση λακίς.	[ἀντ. δ.] 121 125

107 ἀπάτην QP² 108 θηγάς A Trl. 110 ἐπέτεός] trisyllab. 1 cf. Theb. 78 ἀνάσσω Brunck 111 σοῖσσω codd.: corr. Herm.: (παρὰ) Wellauer: παρασινοῖς βροτῶν Seidler ceteris tanquam scholio in textum recepto deletis 112 ἄρκυας ἀνα Herm.: ἀκυστάτα codd. 113 ὄπτε ΣΦ: sed super retina non subter evadendum est: Ag. 359, 1376 115 μόν MQ 121 ἴσεται (M) vel ἴσεται codd.: corr. Burney 125 πίσση λακίς om. M: add. m

Edizione critica

ΑΙΣΧΥΛΟΥ		ΠΕΡΣΑΙ	
ἐχτροίσι πεποθῶς στιφυλαῖς ἐφέταις, χρο- σογόνου γενεᾶς ἰσθίους φῶς.	80	ἐσορᾶν πότιον ἄλοος, πίσινου λεπτοδόμοιο πει- σμαῖσι λα- σπόρος τε μηχαναῖς.	105
κυάνειον δ' ὄμμασι λιύσσων φονίου δέργμα δράκοντος, πολύχειρ και πολυναύτης, λύριον θ' ἄρμα διώκων, ἐπάγει δουρικλύτους ἀν- δράσι τοξόδαμνον Ἀρη.	[στρ. β.] 85	δολόμητην δ' ἀπάταν θεοῦ τίς ἀνηρ θνατός ἀλύξει; τίς δ' κραπιγῶ ποδὶ πηδή- ματος εὐπέτερος ἀνάσσει; φιλόφρων γάρ (ποτι)σφαίνου- σα τὸ πρώτον παραίγει βροτῶν εἰς ἄρκυος Ἄτα, τόθεν οὐκ ἔστιν ὑπὲρ θνα- τῶν ἀλύξαντα φυγεῖν.	93 [μεσσηδ.] 95 110
δόκιμος δ' οὔτις ὑποστάς μεγάλῳ βρέμματι φωτῶν ἐχτροῖς ἔρκεσιν εἰργχειν ἄμαχον κῆμα θαλάσσης· ἀπράσιτος γάρ ὁ Περσῶν στρατὸς ἀλκίφρων τε λαός.	[ἀντ. β.] 90	ταπᾶ μοι μελαγχίτων φρῆν ἀμύσσειται φόβω— ὃ Ἄπερακοῦ στρατεύματος— τοῦδε μὴ πάλαι πύθη- ται, κένανθρον μὲν' ἄστῃ Σουσίδοι·	[στρ. δ.] 116
θεῖθον γὰρ κατὰ Μοῖρ' ἐκράτησεν τὸ παλαιόν, ἐπέακμησέ δὲ Πέρσαις πολύμοις πυργουδαίκοις διέπειν ἱπποχάρμας τε κλύουσι πάλεόν τ' ἀναστάσεις.	101 [στρ. γ.] 95 105	καὶ τὸ Κισσιῶν πόλιον μ' ἀντιδουπον ἤσεται, ὃ δ', τοῦτ' ἔπος γυναικοπλη- θῆς ὁμιλος ἀπύων, βυσοῖσιν δ' ἐν πέπλοισι πέσση λαίσι.	[ἀντ. δ.] 121 125
ἔμαθον δ' εὐρυπόροιο θαλάσσης πολαινομένης πνεύματι λάβρω	[ἀντ. γ.] 110 101		

PQ στυφίλους codd. 79-80 χρυσογόνου memorum ut v.l. (5 καὶ βλάνου) ΣΜΣΦQP; χρυσογόνου codd. 80 ἰσθίων M 81 κυάνειον] κυανῶν Blomfield: cf. Theb. 78 82 φονίου M²: φονίων fere tell. ἄρμα M 83 πολυναύτης A 84 εὐρόων M²Trl.: δασύων vel ἀσύρων tell. 86 ἄρη PVQ; cf. Theb. 45 89 ὄρχοις Φ: cf. 78 et Ag. 44 93-100 pott. vv. 101-114 tralicit O. Müller 99 πάλων δ' fere codd.: corr. Byz. 101 θαλάσσης MΦ πολυνομένη M²V

107 ἀπάτην QP² 108 θητόν A Trl. 110 εὐπέτερος] trisyllab.: cf. Theb. 78 ἀνάσσει Brunck 111 σοῖσσοι codd.: corr. Herm.: (παρὰ) Wellauer: παρασάσεις βροτῶν Seidler ceteris tanquam scholio in textum recepto deletis 112 ἄρκυος ἀνα Herm.: ἀκυστάτα codd. 113 ὄπεκ ΣΦ: sed super retina non subter evadendum est: Ag. 359, 1376 115 μόν MQ 121 ἔσεται (M) vel ἔσεται codd.: corr. Burney 125 πέσση λαίσι om. M: add. π

Edizione critica

ΑΙΣΧΥΛΟΥ		ΠΕΡΣΑΙ	
ἐχνοροῖσι πεποθῶς στιφυλοῖς ἐφέταις, χρυ- σογόνοι γενεᾶς ἰσθίους φῶς.	80	ἐσορᾶν πότιον ἄλοος, πίσνοι λεπτοδόμοις πεί- μμασι λα- σπόροις τε μηχαναῖς.	105
κύνειον δ' ὄμμασι λυίσσον φονίου δέργμα δράκοντος, πολύχειρ και πολυναύτης, λύριον θ' ἄρμα δίκων, ἐπάγει δουρικλήτους ἀν- δράσι τοξόδομον Ἀρη.	[στρ. β.] 85	δολόμενῃ δ' ἀπάταν θεοῦ τίς ἀνηρ θνατός ἀλύξει; τίς ὁ κρασιπῆγ' ποδὶ πηδή- ματος εὐπέτερος ἀνάσσει; φιλόφρων γάρ (ποτι)σθαῖνοι- σα τὸ πρῶτον παραγεί- βροτῶν εἰς ἄρκους ἄτα, τόθεν οὐκ ἔστιν ὑπὲρ θνα- τῶν ἀλύξαντα φυγεῖν.	93 [μισαυδ.] 95 110 100
δόκιμος δ' οὔτις ὑποστάς μεγάλοι βρέμυται φωτῶν ἐχνοροῖς ἔρκεσιν εἰργειν ἄμαχον κῆμα θαλάσσης· ἀπράσιτος γάρ ὁ Περσῶν στρατὸς ἀλκίφρων τε λαός.	[ἀντ. β.] 90	ταυτὰ μοι μελαγχίτων φρήν ἀμύσσειται φόβω— ὁἶ Περσικοῦ στρατεύματος— τοῦδε μὴ πάλαι πύθη- ται, κένανθρον μὲν' ἄστν Σουσιδος·	[στρ. δ.] 116
θεῖθεν γὰρ κατὰ Μοῖρ' ἐκράτησεν τὸ παλαιόν, ἐπέσκηψε δὲ Πέρσας πολύμους πυργουδαίτους διέπειν ἱπποχάρμας τε κλύουσι πάλαιον τ' ἀναστάσεις.	101 [στρ. γ.] 95 105	καὶ τὸ Κισσιῶν πόλιον' ἀντίθουπον ἄσεται ὁἶ, τοῦτ' ἔπος γυναικοπη- θῆς ὁμιλος ἀπύων, βυσσοῖσι δ' ἐν πέπλοισι πύση λαῖσι.	[ἀντ. δ.] 121 125
ἐμβαθὸν δ' εὐρυπτόροιο θαλάσσης πολαινομένης πνεύματι λάβρω	[ἀντ. γ.] 110 101		

PQ στυφίλοισι codd. 79-80 χρυσογόνοις ποσειφάντι ut v.l. (5 καὶ
 βλάνου) ΣΜΣΦQP; χρυσοφόμοι codd. 80 ἰσθίουθον M 81 κύνειον
 κνωσῶν Blomfield; cf. Theb. 78 82 φονίου M; φονίου fere tell.
 ἄρμα M 83 πολυναύτας A 84 εὐνοῖον M Trl.; δεσφίον vel
 ἀσφίον tell. 86 ἄρμη PVQ; cf. Theb. 45 89 ὄρχοις Φ; cf. 78 et
 Ag. 44 91-100 PBM vs. 101-114 Trileit O. Müller 99 ποδίων
 δ' fere codd.; corr. Byz. 101 θαλάσσης M Φ πολυνομένη M Φ

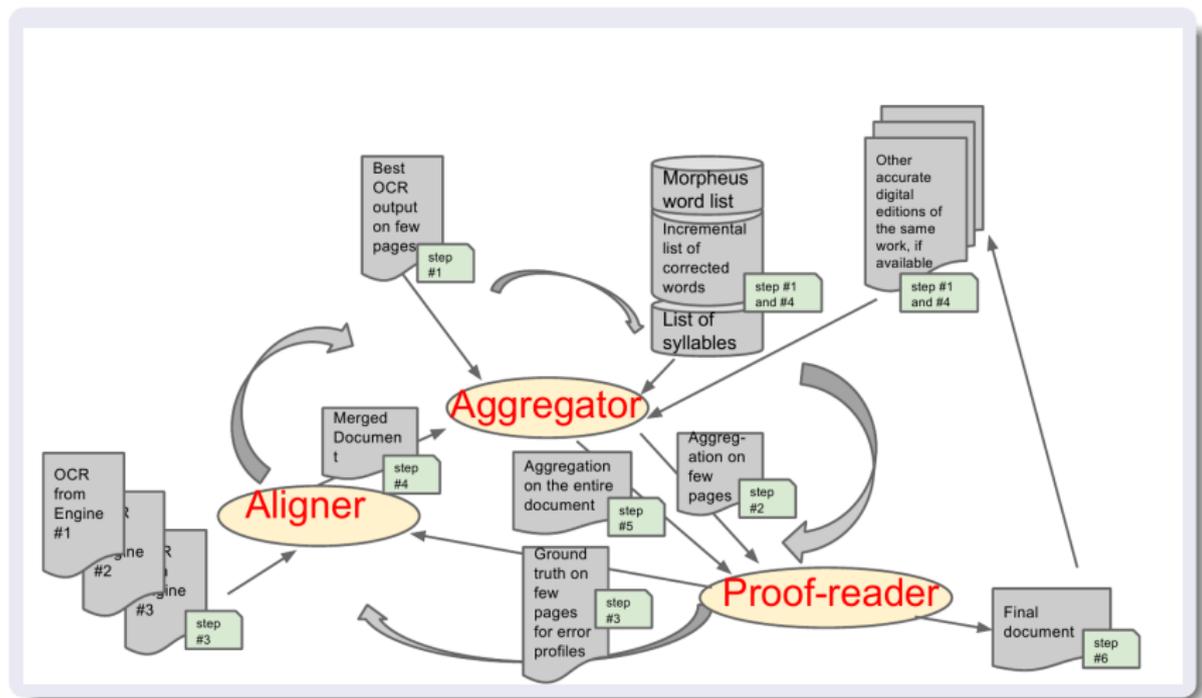
107 ἀπάτην QP² 108 θητότ' A Trl. 110 εὐπέτερος] trisyllab.;
 cf. Theb. 78 ἀνίσσων Brunck 111 σοῖσιν codd.; corr. Herm.;
 (παρὰ) Wellauer: παρασῶνιες Brunck Seidler ceteris tanquam scholio in
 textum recepto deletis 112 ἄρκουσι ἀνα Herm.; ἀκρότατοι codd.
 113 ὄπτε ΣΦ; sed super retia non subter evadendum est: Ag. 359, 1376
 115 μόν MQ 121 ἄσεται (M) vel ἄσεται codd.; corr. Burney
 125 πύση λαῖσι om. M; add. π

Peculiarità dell'OCR storico

Peculiarità dell'OCR applicato a documenti di interesse storico

- Qualità variabile della carta e degli inchiostri (materie prime usate)
- Qualità variabile dell'impressione a stampa (tecnologia impiegata)
- Stato di conservazione (fattore temporale)
- Varietà linguistiche assenti nei dizionari di riferimento (variabili diacroniche, diatopiche, diastratiche, diafasiche, diamesiche)

OCR Work-flow



Overview

- 1 **Trattamento delle immagini**
- 2 OCR e post-processing
- 3 Correzione collaborativa e cooperativa

I repertori di immagini online

<https://archive.org>

“Internet Archive is a non-profit library of millions of free books, movies, software, music, websites, and more.”

<https://www.hathitrust.org>

“HathiTrust is a partnership of academic and research institutions, offering a collection of millions of titles digitized from libraries around the world.”

Altre risorse

- <https://www.bsb-muenchen.de>
- <https://books.google.com>
- ...

Edizioni ed esemplari

- Online non solo è possibile trovare molteplici edizioni della stessa opera, ma anche molteplici esemplari della stessa edizione
- L'applicazione dell'OCR a più esemplari della stessa edizione e l'allineamento dei risultati possono migliorare l'accuratezza della digitalizzazione
- È buona norma catalogare gli esemplari disponibili online, possibilmente secondo il modello denominato *Functional Requirements for Bibliographic Records* (FRBR) raccomandato dall'*International Federation of Library Associations* (IFLA)

La scansione

In assenza di immagini di alta qualità disponibili online, è necessario procedere alla scansione delle proprie immagini. Se i materiali non sono delicati (edizioni antiche, preziose, facili da rovinare), un'economico scanner piatto può essere sufficiente

DPI

- La risoluzione delle immagini acquisite dipende dal miglior compromesso fra qualità e costo (costo per gli strumenti di acquisizione e per lo *storage*)
- Attualmente, 600 punti per pollice (*dots per inch*: DPI) sono considerati sufficienti per ottenere risultati ottimali con l'OCR
- È buona norma salvare i *master files* con le immagini originali al massimo della risoluzione (anche molto superiore a 600 DPI) e applicare le necessarie trasformazioni a copie delle immagini, eventualmente a risoluzione inferiore

ScanTailor

<http://scantailor.org>

“Scan Tailor is an interactive [...] tool for scanned pages. It performs operations such as page splitting, deskewing, adding/removing borders, and others. You give it raw scans, and you get pages ready to be printed or assembled into a PDF or DJVU file. Scanning, optical character recognition, and assembling multi-page documents are out of scope of this project.”



ScanTailor GUI & CLI

ScanTailor può essere usato sia interattivamente, attraverso l'interfaccia grafica, sia tramite linea di comando, per progetti massivi

Page Splitting, Content Selection and Margins

The screenshot displays the Scan Tailor 0.9.11.1 [64bit] application window. On the left, a sidebar contains a list of steps: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content, 5 Margins, and 6 Output. Step 2 is currently selected. Below this list, the 'Page Layout' section shows three icons representing different page orientations (left-to-right, right-to-left, and two-page spread) with the text 'Auto detected' and a 'Change ...' button. The 'Split Line' section has 'Auto' and 'Manual' buttons. The main window shows a preview of a scanned document with a vertical blue line indicating the split position between two pages. On the right, a vertical stack of thumbnails shows the resulting pages, with the first page labeled 'i0038.png' and the second page labeled 'i0039.png'. A large question mark is overlaid on the second page thumbnail. At the bottom of the right panel, the text 'Natural order' is visible.

Use the mouse wheel or +/- to zoom. When zoomed, dragging is possible.

Page Splitting, Content Selection and Margins

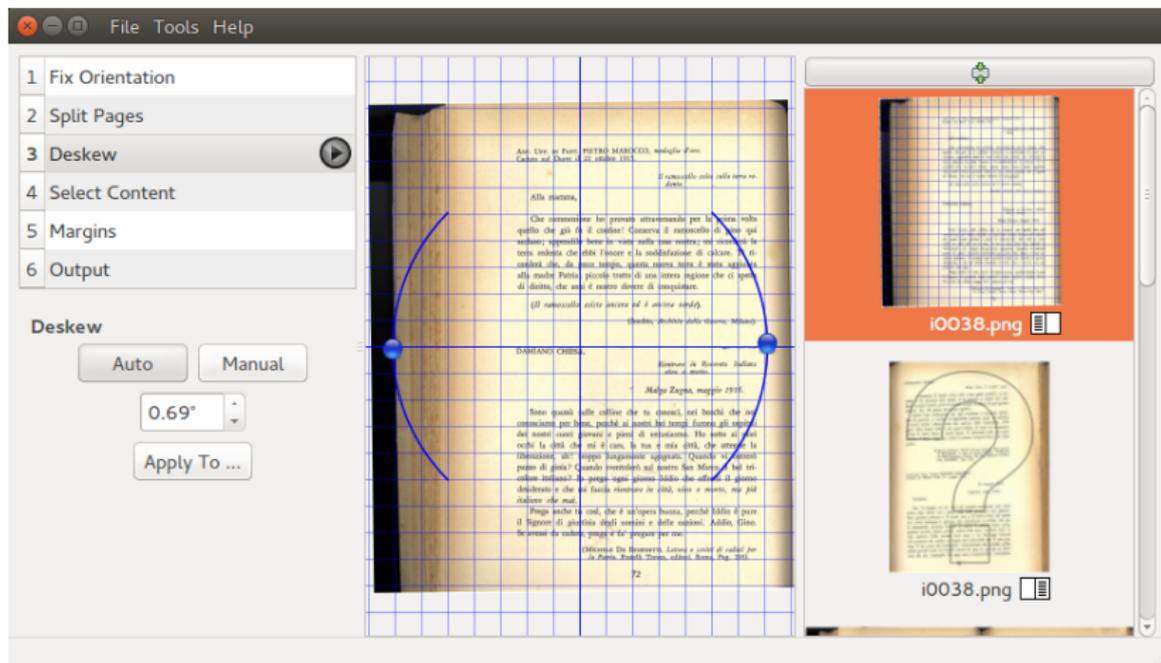
The screenshot displays the Scan Tailor 0.9.11.1 [64bit] application window. On the left, a sidebar contains a menu with six items: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content (highlighted in orange), 5 Margins, and 6 Output. Below the menu is a 'Content Box' with 'Auto' and 'Manual' buttons, and a 'Scope' section with an 'Apply to ...' button. The main workspace shows a scanned page with a blue selection box around the text. The text includes: 'Car. Cav. di Feltre PIETRO MEMMO, consigliere della Camera di Commercio di Treviso il 20 ottobre 1915.', 'Alla mamma,', 'Che commossa ho provato attraversando per la prima volta quello che già fu il cimitero! Conserva il ricordo di più qui vicino; appesantilo bene in vista della tua madre; mi ricorderei le tue labbra che dala' l'emozione e la soddisfazione di colare. Ti si corderà che, da poco tempo, questa nuova terra è stata aggiunta alla madre Patria; piccolo tratto di una istruita ragione che si spazia al di là, che non è nuova diversa di occupazione.', 'Il cimitero delle anime ed il nostro mondo.', '(Dedica, Archivi della Guerra, Milano)', 'RAMANDO CHISA.', 'Bianco in Bianco Italiano', 'Maja Zagna, maggio 1915.', 'Sono quasi sulle colline che tu amasti, nei boschi che noi amavamo per bene, perché ai nostri bei tempi furono gli ospitali dei nostri cuori giovani e pieni di entusiasmo. Ho visto ai miei piedi la città che noi è cara, la tua e mia città, che attende la Riformazione, ah! sempre lungamente sognata. Quando vi tornate piano di gioia? Quando ritornerete nel nostro suo Milano il bel mattino italiano? Ho pregato ogni giorno Milano che offerti il giorno desiderato e che mi faccia ritrovare in città, viva o morta, ma più felice che mai.', 'Prega anche tu così, che è un'opera buona, perché Milano è parte il Signore di giustizia degli uomini e delle machine. Addio, Gina. Ho amato da vicino, pregò e le' pregare per noi.', 'Ottaviano De Bonaventis. Lettera a cavigli di calce per la Anna. Firenze Treves, editore, Roma, Pag. 1013', '32'. On the right, a preview pane shows two thumbnails of the scanned page with selection boxes, labeled 'i0038.png'. Below the thumbnails is a 'Natural order' label. The bottom of the window features a navigation bar with various icons for navigation and search.

Page Splitting, Content Selection and Margins

The screenshot displays the Scan Tailor 0.9.11.1 [64bit] application window. On the left, a sidebar contains a menu with '4 Select Content', '5 Margins' (highlighted), and '6 Output'. Below the menu, the 'Margins' section is active, showing 'Millimeters (mm)' and input fields for Top (5.0), Bottom (5.0), Left (10.0), and Right (10.0). An 'Apply To ...' button is located below the input fields. The 'Alignment' section below has a checked 'Match size with other pages' option and several alignment icons. The main workspace shows a scanned page of text with a pink rectangular selection box around the content. The text on the page includes: 'Ass. Uff. di FANI. PIETRO MARCOCCO, *malattia d'ero.* Caduto sul Duver il 22 ottobre 1915.', 'Il ramoscello colto sulla terra redenta.', 'Alla mamma,', 'Che commoione ho provato attraversando per la prima volta quello che già fu il confine! Conserva il ramoscello di pino qui accluso; appendilo bene in vista nella casa nostra; mi ricorderà la terra redenta che ebbi l'onore e la soddisfazione di calcare. Ti ricorderà che, da poco tempo, questa nuova terra è stata aggiunta alla madre Patria; piccolo tratto di una intera regione che ci spetta di diritto, che anzi è nostro dovere di conquistare.', '(Il ramoscello esiste ancora ed è ancora verde). (Inedito, Archivio della Guerra, Milano). DAMIANO GHISA. Rientrare in Rovereto Italiano vivo o morto. Malga Zugna, maggio 1916. Sono quasi sulle colline che tu conosci, nei boschi che noi conosciamo per bene, perché ai nostri bei tempi furono gli ospitali; lei nostri cuori giovani e pieni di entusiasmo. Ho sotto ai miei occhi la città che mi è cara, la tua e mia città, che attende la liberazione, ah! troppo lungamente agognata. Quando vi entrerò pazzo di gioia? Quando sventolerò sul nostro San Marco il bel tricolore italiano? Io prego ogni giorno Iddio che affretti il giorno desiderato e che mi faccia rientrare in città, vivo o morto, ma più italiano che mai. Pregha anche tu così, che è un'opera buona, perché Iddio è pure il Signore di giustizia degli uomini e delle nazioni. Addio, Gino. Se avessi da cadere, prega e fa' pregare per me. (MICHELE DE BERNIETTI, *Lettere e scritti di caduti per la Patria*. Fratelli Treves, editori. Roma, Pag. 291). 72

On the right, two preview windows are visible. The top one shows a thumbnail of the selected content with the filename 'i0038.png'. The bottom one shows a thumbnail of the page with a large question mark overlaid, also labeled 'i0038.png'. At the bottom right of the application window, there are navigation icons for search, zoom, and refresh.

Deskewing



Dewarping

The screenshot displays the Scan Tailor 0.9.11.1 [64bit] application window. On the left, a sidebar contains a menu with six items: 1 Fix Orientation, 2 Split Pages, 3 Deskew, 4 Select Content, 5 Margins, and 6 Output. Below the menu, the 'Output Resolution (DPI)' is set to 600, with a 'Change ...' button. The 'Dewarping' section is set to 'Manual', also with a 'Change ...' button. A 'Depth perception' slider is positioned at the far left, with an 'Apply To ...' button below it. The central workspace shows a scanned page with a blue grid overlay that is warped to follow the curve of the page. On the right, a vertical toolbar includes buttons for 'Output', 'Picture Zones', 'Fill Zones', 'Dewarping', and 'Despeckling'. To the right of the toolbar, a preview area shows the original scanned page (top) and the result after dewarping (bottom), both labeled 'i0038.png' with a thumbnail icon.

Binarization

The screenshot displays the Scan Tailor 0.9.11.1 [64bit] interface. On the left, a sidebar contains a list of steps: 4 Select Content, 5 Margins, and 6 Output. Below this, the 'Output Resolution (DPI)' is set to 600. The 'Mode' is set to 'Black and White'. A 'Thinner' / 'Thicker' slider is positioned at the center. The 'Dewarping' option is currently 'Off'. The 'Despeckling' section at the bottom has three icons. The main workspace shows a document page with the following text:

DAMIANO CHIESA

Veder libero il proprio paese.

... il pensiero di essere vicini così e non poter andarvi, ci rattrista, ci fa ricordare altri tempi; ci fa pensare ai nostri cari confinati in paesi lontani, privi di notizie, sempre in attesa di quel giorno solenne. Ah, che giorno dev'essere quello!...

Anche a noi combattenti, che ogni momento ci troviamo circondati da pericoli, ci sembra un'ingiustizia enorme, quasi un'infamia il dover morire adesso, oltre che nell'ora della redenzione, sulle porte della nostra città. A noi quasi sembra di avere un sacrosanto diritto di veder libero il nostro paese, di chiamarlo tutto col nome fatidico d'Italia; dopo poi, anche il morire c'importa fino a un certo punto.

(Legione Trentina. *Memorie ed eroi trentini, della guerra di Redenzione*, a cura di Oreste Ferrari. Prefazione di Carlo Dolcini. Trento, Tip. Ed. Mutilati ed Invalidi, MCMXXV, Pag. 318).

CAPITANO BRIGLI ALPES BESOZZI MARTINO.
Caduto sul Monte Cukla 711 maggio 1916.

24 maggio 1915.

L'apertura delle ostilità.

Carissimi,

Ieri, 25 maggio, ad ore 18,55 gli austriaci salutavano con i loro primi colpi (forse non i primi colpi della guerra) il mio plotone. Due granate cadevano a 50 metri una, a 10 metri l'altra dal punto ore avevo radunato il plotone, per comunicare ai soldati che per la mezzanotte avevamo l'ordine di aprire le ostilità. Della prima granata raccolta subito, perché caduta sulla neve, ho fatto dono al mio capitano. Della seconda farò dono a voi. Ho oggi ricevuto alla presenza dei soldati un elogio dal Colonnello per il mio plotone. E ho avuto dal Colonnello l'assicurazione che avrebbe scritto subito perché Carlo (il fratello morto lui pure in guerra) sia destinato alla mia compagnia. Da oggi sono completamente comandante

73

On the right side of the interface, there are vertical labels for 'Output', 'Picture Zones', 'Fill Zones', 'Dewarping', and 'Despeckling'. The 'Output' window shows a preview of the binarized page, with a large orange rectangular area covering the central text. Below the preview, the filename 'i0038.png' is displayed. At the bottom right of the interface, there are navigation icons for search, zoom, and refresh.

Despeckling

spettivo e prospe
questa collocazione
altri e piú crudi c
esiti affiorassero
politica che nur

Despeckling

spettivo e prospe
questa collocazione
altri e piú crudi c
esiti affiorassero
politica che n

Ulteriori operazioni

original	ἀπογεύονται	ἀπογεύομαι	CHARSEQ
threshold	ἀπογεύονται	α3πογεύομαι	CHARSEQ
scale	ἀπογεύονται	δπογεδονται	GREEKSEQ
erosion	ἀπογεύονται	ùno211εύουιαι	CHARSEQ
blur	ἀπογεύονται	δπογεύοπαι	GREEKSEQ
dilation	ἀπογεύονται	ἀπο."εύουιαι	CHARSEQ
erosion	ἀπογεύονται	ἀπο'ι'εύονται	GREEKSEQ
erosion	ἀπογεύονται	ἀπογεύονται	WORD

Creare i propri script

```
import numpy as np
import cv2, sys, os
in_img=sys.argv[1]
out_img=in_img[0:-4]+"_clean.png"
im=cv2.imread(in_img)
gray = cv2.cvtColor(im,cv2.COLOR_BGR2GRAY)
grayout=gray.copy()
blur = cv2.GaussianBlur(gray,(5,5),1)
thresh = cv2.adaptiveThreshold(blur,255,1,1,11,2)
img,contours,hierchy = cv2.findContours(thresh,cv2.RETR_TREE,cv2.CHAIN_APPROX_NONE)
wcnt = 0
for item in contours:
    area =cv2.contourArea(item)
    [x,y,w,h] = cv2.boundingRect(item)
    if area<sys.argv[2]:
        cv2.drawContours(grayout,contours,wcnt,255,-1)
        wcnt = wcnt + 1
grayout=cv2.GaussianBlur(grayout,(5,5),1)
cv2.imwrite(out_img,grayout)
```

Overview

- 1 Trattamento delle immagini
- 2 OCR e post-processing
- 3 Correzione collaborativa e cooperativa

Abbyy Fine Reader

<https://www.abbyy.com>

- FineReader è una delle applicazioni commerciali dalle prestazioni più alte per l'OCR.
- FineReader è in grado di compiere complesse analisi del *layout* e di riconoscere testi multilingui.
- È possibile addestrare FineReader a riconoscere nuovi caratteri, associando l'immagine dei glifi ai corrispondenti codici Unicode.

Ideatech Anagnostis

<http://anagnostis.soft112.com>

- Anagnostis è un'applicazione commerciale capace di riconoscere, anche senza addestramento, il Greco antico. L'addestramento (training) può migliorare ulteriormente le prestazioni.
- Spiriti e accenti sono trattati separatamente dal corpo del carattere, migliorando la precisione del sistema di riconoscimento.

Tesseract

<https://github.com/tesseract-ocr>

- Tesseract è attualmente una delle applicazioni *open source* per l'OCR che danno risultati più accurati.
- Tesseract necessita un addestramento *ad hoc* per riconoscere il Greco politonico (o qualsiasi nuovo *set* di caratteri). Il riconoscimento di *set* misti di caratteri dà risultati accettabili.
- Il formato dell'*output* è solo testo oppure xhtml arricchito con un *microformat* (hOCR) che registra le posizioni delle parole (o anche dei singoli caratteri) sull'immagine della pagina.

OCRopus / ocropy

<https://github.com/tmbdev/ocropy>

- Anche OCRopus è *open source* e dà prestazioni molto alte.
- OCRopus è in continua evoluzione. Attualmente è sviluppato in python (ocropy).
- OCRopus utilizza reti neurali.

Training

mordeo nō momordi cum o:fed cū e memordi dicebant.Q.Ennius in fatyris meū

rmordeo nō momordi cum o:fed cū e memordi dicebant.Q.Ennius in fatyris meū

rmordeo nom momordi cum o:sed cum e memordi dicebant.Q.Ennius in satyris meum

(x) 269: tort-016/010024.bin.png

inqt non est:at si me canis memorderit.Laberius in gallis de integro patrimonio

inqt non est:at si me canis memorderit.Laberius in gallis de integro patrimonio

inqt non est:at si me canis memorderit.Laberius in gallis de integro patrimonio

(x) 270: tort-016/010025.bin.png

meo centū milia nūmū memordi.Idem.P.Nigidius.Idem Plautus diuerfis ī locis

meo centū milia nūmū memordi.Idem.P.Nigidius.Idem Plautus diuerfis ī locis

meo centum milia nummum memordi Idem PNioidius Idem Plautus diuersis im locis

Gamera / Rigaudon

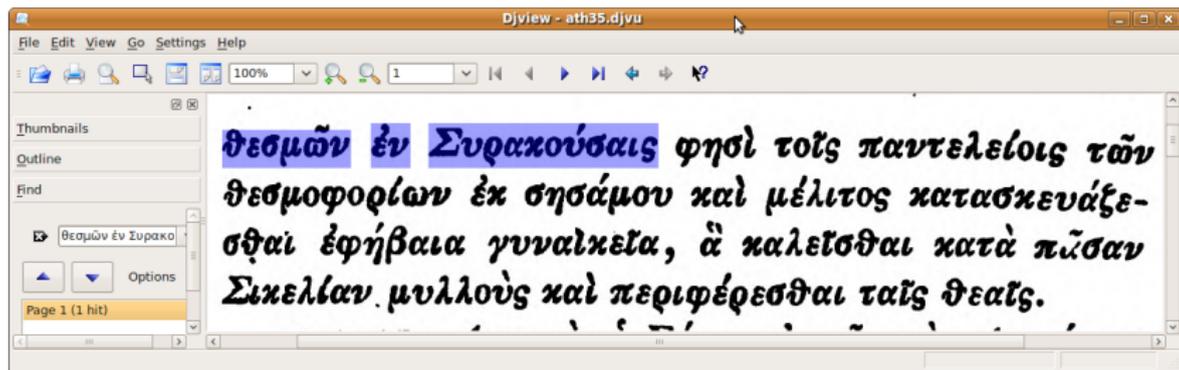
<http://gamera.informatik.hsnr.de/addons/ocr4gamera>

- Anche Gamera, come Tesseract e OCRopus, è *open source* e dà prestazioni molto alte.
- Bruce Robertson ha ottimizzato Gamera e ha creato dei moduli *ad hoc* per il Greco: *Rigaudon* (<https://github.com/brobertson/rigaudon>).

Valutazione dell'accuratezza

$$\textit{accuratezza} = \frac{\textit{corrispondenze}}{\textit{corrispondenze} + \textit{substituzioni} + \textit{inserzioni} + \textit{cancellazioni}}$$

Mappatura del testo sull'immagine



hOCR

```
<p class='ocr_par' id='par_1_7' lang='fra' title="bbox 797 1884 3144 2194">  
  <span class='ocr_line' id='line_1_8' title="bbox 802 1884 3144 2032; baseline -0 -27; x_size 114;  
    x_descenders 31; x_ascenders 28">  
    <span class='ocrx_word' id='word_1_49' title='bbox 802 1884 893 2005; x_wconf 83'>Il</span>  
    <span class='ocrx_word' id='word_1_50' title='bbox 940 1950 1043 2006; x_wconf 91'>en</span>  
    <span class='ocrx_word' id='word_1_51' title='bbox 1083 1926 1360 2032; x_wconf 69'>prend</span>  
    <span class='ocrx_word' id='word_1_52' title='bbox 1408 1952 1529 2007; x_wconf 88'>un</span>  
    <span class='ocrx_word' id='word_1_53' title='bbox 1569 1929 1796 2031; x_wconf 59'>pour</span>  
    <span class='ocrx_word' id='word_1_54' title='bbox 1845 1917 2001 2016; x_wconf 89'>soi,</span>  
    <span class='ocrx_word' id='word_1_55' title='bbox 2056 1917 2259 1998; x_wconf 89'>dont</span>  
    <span class='ocrx_word' id='word_1_56' title='bbox 2308 1916 2392 1999; x_wconf 90'>la</span>  
    <span class='ocrx_word' id='word_1_57' title='bbox 2444 1943 2684 1999; x_wconf 90'>corne</span>  
  </span>  
</p>
```

Allineamento

Esempio di *progressive multiple sequence alignment*

ιερᾶς Μέμφιδος ἄρχων[¶]μέγας Ἀρ^οσά^ομης, τὰς τ^ο ὠγγίους[¶]ἰθήβας ἐφέπων Ἀριόμαρδος,[¶]
ιερᾶς Μέμφιδος ἄρχων[¶]μέγας Ἀρ^οσά^ομης, τὰς τ^ο ὠγγίους[¶]ἰθήβας ἐφέπων Ἀριόμαρδος,[¶]
1 ρᾶς Μέμφιδος ἄρχων[¶]μέγας Ἀρ εἰσ^ομης, τὰς π^ο ὠγιέους[¶]ἰθήβας φσέκων Ἀριόμαρδος,[¶]

καὶ ἐλειοβάται ναῶν ἐρέται, [^]0[¶]ιδεινοὶ πληθός τ^ο ἀνάριθμοι. [¶]ἄβροδιαίτων δ^ο ἐπεται
καὶ ἐλειοβάται ναῶν ἐρέται, ^ο0[¶]ιδεινοὶ πληθός τ^ο ἀνάριθμοι. [¶]ἄβροδιαίτων δ^ο ἐπεται
καὶ ^ολειοβάται ναῶν ἰρέται, 40[¶]ιδεινοὶ πληθός τ^ο ἀνάριθμοι. [¶]ἄβροδιαίτων δ^ο ἐκεται

[^]Αυδῶν[¶]ἰόχλος, οἷτ^ο ἐπίπαν ἠπειρογενῆς^ο [¶]κατέχουσιν ἔ^οθνος, τους Μιτρογαθῆς[¶]Ἄρκτ
^οΛυδῶν[¶]ἰόχλος, οἷτ^ο ἐπίπαν ἠπειρογενῆς^ο [¶]κατέχουσιν ἔ^οθνος, τοὺς Μιτρογαθῆς[¶]Ἄρκτ
^οΛυδῶν[¶]ἰόχιος, οἷτ^ο ἐπίπα^ο ἠπε ρογενῆς^ο [¶]κατέχουσιν ε^οἄνος, τοὺς Μιτροκα^οης[¶]Ἄρ^ο

εὐς τ^ο αγαθός, βασιλῆς δίσποι, [^]ἴχαι πολύχρσοι ^οΣάρδεις ἐπόχους[¶]πολλοῖς ἄρμασιν
εὐς τ^ο ἀγσθός, βασιλῆς δίσποι, ^οἴχαι πολύχρσοι ^οΣάρδεις ἐπόχους[¶]πολλοῖς ἄρμασιν
εὐς τ^ο αγααός, βασιλῆς δ^οπο^ο α^οἴαα^ο πολ^οευσσι -σάρδαις ἐπόχους[¶]πολλοῖς ἄρμασιν

ἐ^οξορμῶσιν, [¶]ἰδῖρρυμά τε καὶ τρίρρυμα τέλη[¶]φοβερὰν ὀ^οψιν προσιδέσθαι. [¶]ἰστεῦται δ^ο
ἐ^οξορμῶσιν, [¶]ἰδῖρρυμά τε καὶ τρίρρυμα τέλη[¶]φοβερὰν ὀ^οψιν προσιδέσθαι. [¶]ἰστεῦται δ^ο
ἐ^οζορμῶσιν, [¶]ἰδῖρρυμά τι κα^ο πρέρρυμα τέλη[¶]φοβερὰν ὀ^οψιν προσιδέσθαι. [¶]ἰστεύτσει δ^ο

Selezione

ἄλλος δ' ἐκείνου παῖς τόδ' ἔργον ἤνυσεν.

ἄ	λ	λ	ο	ς	δ	'	ε	κ	ε	ί	ν	ο	υ	-	π	α	ῖ	ς	τ	ό	δ	'	έ	-	ρ	γ	ο	ν	η	ν	υ	σ	ε	ν	.	FineReader			
																																						OCRopus	
ἄ	λ	λ	ο	ς	δ	'	έ	κ	ε	ί	ν	ο	υ	*	π	α	ῖ	ς	τ	ό	δ	'	έ	'	ρ	γ	ο	ν	η	ν	υ	σ	ε	ν	.	Anagnostis			
																																						Result	
;	λ	λ	ο	ς	-	ό	έ	χ	ε	;	τ	ο	υ	-	-	κ	α	-	ς	τ	ό	δ	-	-	-	ρ	γ	ο	>	η	ν	υ	σ	ι	ν	.			
ἄ	λ	λ	ο	ς	δ	'	έ	κ	ε	ί	ν	ο	υ	-	π	α	ῖ	ς	τ	ό	δ	'	έ	-	ρ	γ	ο	ν	η	ν	υ	σ	ε	ν	.				

Spell-checking

Output di FineReader	RegEx per tutte le applic. OCR	Suggerimenti dello spell-checker	Risultati
ἐξερήμωσεν ωπασεν εν' επάσης εὐθυνητήριον πρώτος Κύρος εθηκε Δυδῶν λάδον ἦλασεν ευφρων	ἐξερήε?[μι]ωσεν [ωοῶ]π[αο]σ[εό]ν [εἶ]ν' ε?ά?πάσης [εἶ][ύυ]θυνητ[ήη]ριον πρ[ῶῶ]τος [ΚΧΗ][ύῦ]ρος [εἶ]θηκε [ΔΛ]υδῶν λ[αά][όο]ν [ήή]λασ[ετ]?ν ε?ι?[υδ]φρωο?ν	ἐξερήμωσε, ἐξερήμωσέ, ἐξερήμωσεν ῶπασεν, ῶπασέν, σπασέν έν, έν' ... ἔν' (34. elemento) πάσης, πάσης ... ἀπάσης (11. elemento) εὐθυνητήριον, εὐθυνητήριόν, εὐθυνητήρι πρῶτος, πρῶτός, πρωτὸς Κῦρος, Κῦρός, Κύπρος ἔθηκε, ἔθεικέ, θήκε Δυῶν, Διδῶν ... Λυδῶν (6. item) λαδόν, λαόν, Λαίον ἦλασεν, ἦλασέν, ἦασεν εὐφρων, Εὐφρων, εὐφρων (corretto)	ἐξερήμωσεν ῶπασεν έν' ἀπάσης εὐθυνητήριον πρῶτος Κῦρος ἔθηκε Λυδῶν λαδόν ἦλασεν ευφρων

Suggerimenti correttamente accettati

Output di FineReader

ωπασεν

RegEx per tutte le applic. OCR

[ωοώ]π[αο]σ[έο]ν

Suggerimenti dello spell-checker

ώπασεν, ώπασέν, σπάσεν

Risultato

ώπασεν

Suggerimenti correttamente rifiutati

Output di FineReader

έξερήμωσεν

RegEx per tutte le applic. OCR

έξερήέ?[μι]ωσεν

Suggerimenti dello spell-checker

έξερήμωσε, έξερήμωσέ, έξηρήμωσεν

Risultato

έξερήμωσεν

Suggerimenti scorrettamente rifiutati

Output di FineReader

ευφρων

RegEx per tutte le applic. OCR

$\epsilon?ι?[υ\delta]\phi\rho\rho\omega\theta?v$

Suggerimenti dello spell-checker

ἐύφρων, Εϋφρων, **εϋφρων** (corretto)

Risultato

ευφρων

Overview

- 1 Trattamento delle immagini
- 2 OCR e post-processing
- 3 **Correzione collaborativa e cooperativa**

Collaborativo / Cooperativo

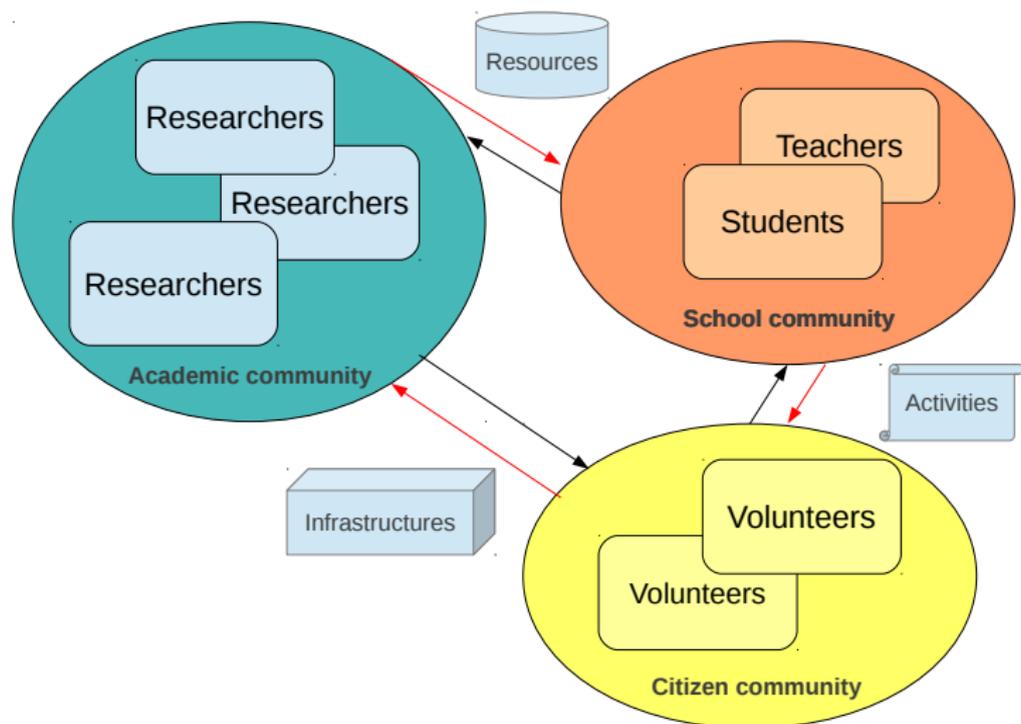
Collaborazione

La collaborazione implica l'interazione di più soggetti o di più gruppi con un obiettivo comune. Il prodotto ottenuto è frutto di negoziazioni fra le parti e costituisce il raggiungimento di tale obiettivo.

Cooperazione

La cooperazione implica la possibilità di suddividere un obiettivo in parti modulari, in modo che i sottoprodotti siano autonomi e riusabili da terze parti per scopi non necessariamente previsti dagli autori originari

Coinvolgere comunità differenti



WIKISOURCE

English

The Free Library

765,000+ pages

Français

La bibliothèque libre

1,209,000+ pages

Русский

Свободная библиотека

293,000+ статей

Deutsch

Die freie Quellensammlung

349,000+ Seiten

Português

A biblioteca livre

28,000+ páginas



Español

La biblioteca libre

102,000+ páginas

Italiano

La biblioteca libera

106,000+ pagine

עברית

הספרייה החופשית

מאמרים 141,000+

Polski

wolna biblioteka

العربية

المكتبة الحرة

Lo studio del greco nell'Europa del XV secolo

Progetto Marie Curie di Paola Tomè

<http://greek15century.mml.ox.ac.uk>

Greek Studies in XVth Century Europe

HOMEPAGE ABOUT RESOURCES HIGH SCHOOLS INVOLVEMENT PEOPLE NEWS AND EVENTS CONTACTS MORE



Erasmus

"Greek Studies in 15th Century Europe" is a Marie Curie Individual research project (2015-17) held by Paola Tomè and based at the Medieval and Modern Languages Faculty, University of Oxford, under the supervision of Martin McLaughlin and Nigel Wilson. The purpose is to investigate the crucial role played by the return of knowledge of Greek in the transformation of European culture, both through the translation of texts, and through the direct study of the language.



Search...

RECENT NEWS AND EVENTS

WORKSHOP IN TURIN, BNU, 29 - 30 JUNE 2017
"Biblioteche private e produzione di libri manoscritti greci a Venezia nel Cinquecento"
Biblioteche private e produzione di libri manoscritti greci a Venezia nel Cinquecento
Giovedì 29 giugno [...]

Orthographia di Giovanni Tortelli



Sciogliere o non sciogliere le abbreviazioni?

- In fase di training i ricercatori hanno trascritto accuratamente, carattere per carattere, il testo rispettando le abbreviazioni originarie, con l'impiego dei caratteri Unicode raccomandati dalla *Medieval Unicode Font Initiative* (MUFI)
- In fase di correzione cooperativa gli studenti hanno sciolto le abbreviazioni

CoPhi Proof-reader per il greco antico

The screenshot shows a web browser window with the URL `localhost:3000/CoPhiProofReader`. The page title is "Euclides, Opera1" and there is a "Save" button. The main content is a Greek text document with line numbers 133, 208, and 344 on the left. The text is as follows:

133 τέρον πλευράν, οὕτως ἢ ἕτερά του δευτέρου πλευρά
208 τέρου πλευράν, οὕτως ἢ ἕτερά του δευτέρου πλευρά
344 10 πρὸς ἦν ἢ λοιπὴ του πρώτου λόγον ἔχει δεδομέναι.
10 πρὸς ἦν ἢ λοιπὴ του πρώτου λόγον ἔχει δεδομένον.

δύο γὰρ παραλληλόγραμμα τὰ AB, EH πρὸς ἄλληλα
δύο γὰρ παραλληλόγραμμα τὰ AB, EH πρὸς ἄλληλα

λόγον ἔχτω δεδομένον ἦτοι ἐν ἴσας γωνίας ἢ ἐν
λόγον ἔχτω δεδομένον ἦτοι ἐν ἴσας γωνίας ἢ ἐν

ἀνίσαις μὲν, δεδομέναις δέ, ταῖς πρὸς τοῖς Γ, Ζ· λέγω,
ἀνίσαις μὲν, δεδομέναις δέ, ταῖς πρὸς τοῖς Γ, Ζ· λέγω, ὅτι

ὅτι ἐστὶν ὡς ἢ GB πρὸς τὴν ZH, οὕτως ἢ EZ πρὸς
ὅτι ἐστὶν ὡς ἢ GB πρὸς τὴν ZH, οὕτως ἢ EZ πρὸς

15 ἦν ἢ ΑΓ λόγον ἔχει δεδομένον.
15 ἦν ἢ ΑΓ λόγον ἔχει δεδομένον.

τὸ γὰρ AB τῷ EH ἦτοι ἰσογώνιον ἐστὶν ἢ οὐ.
τὸ γὰρ AB τῷ EH ἦτοι ἰσογώνιον ἐστὶν ἢ οὐ.

Annotations on the right side of the text:

- Wrong accents and breathing marks:** Points to `οὕτως` and `ἕτερά` in the first two lines.
- Self-corrections:** Points to `ἴσας` in the third line.
- Errors:** Points to `οὕτως` in the fifth line.

Lista dei suggerimenti

Euclides, Opera1 Save

133
208
344

τέρου πλευράν, οὕτως ἢ ἑτέρα τοῦ δευτέρου πλευρᾶ
τέρου πλευράν, οὕτως ἢ ἑτέρα τοῦ δευτέρου πλευρᾶ

10 πρὸς ἦν ἢ λοιπῇ τοῦ πρώτου λόγον ἔχει δεδομένον.
10 πρὸς ἦν ἢ λοιπῇ τοῦ πρώτου λόγον ἔχει δεδομένον.

δύο γὰρ παραλληλόγραμμα τὰ AB, EH πρὸς ἄλληλα
δύο γὰρ παραλληλόγραμμα τα AB, EH πρὸς ἄλληλα

λόγον ἔχτω δεδομένον ἦτοι ἐν ἴσας ἢ ἐν
λόγον ἔχτω δεδομένον ἦτοι ἐν ἴσας γ ἢ ἐν

ἀνίστοις μὲν, δεδομέναις δέ, ταῖς πρὸς τὴν
ἀνίστοις μὲν, δεδομέναις δέ, ταῖς πρὸς τὴν πρᾶως λέγω,
πρᾶως λέγω,

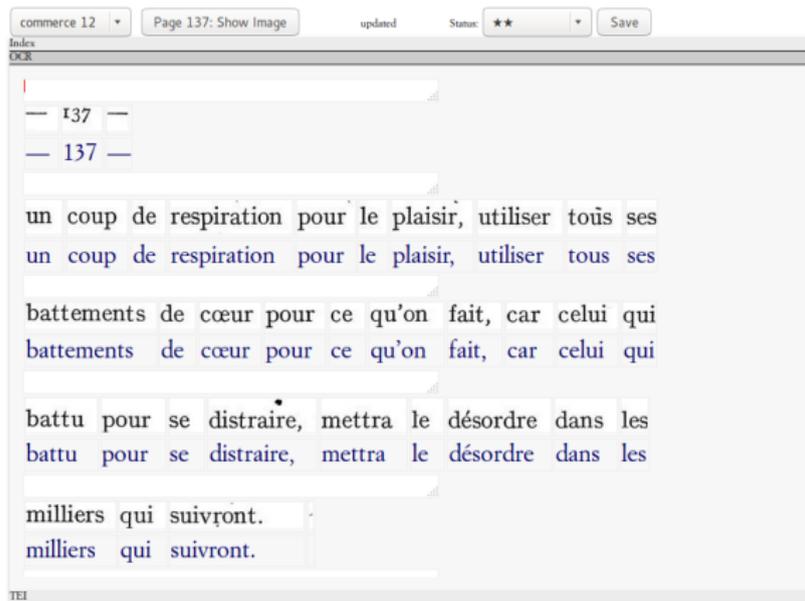
ὅτι ἐστὶν ὡς ἢ GB πρὸς τὴν ZH , οὕτως ἢ EZ πρὸς
ὅτι ἐστὶν ὡς ἢ GB πρὸς τὴν ZH , οὕτως ἢ EZ πρὸς

15 ἦν ἢ AG λόγον ἔχει δεδομένον.
15 ἦν ἢ AG λόγον ἔχει δεδομένον.

τὸ γὰρ AB τῷ EH ἦτοι ἰσογωνίων ἐστιν ἢ οὐ.
τὸ γὰρ AB τῷ EH ἦτοι ἰσογωνίων ἐστιν ἢ οὐ.

CoPhi Proof-reader per Commerce Online

Progetto coordinato da Antonietta Sanna, Università di Pisa
Digitalizzazione di 29 volumi della rivista letteraria *Commerce*



Reintegrare la formattazione

```
TEI
(re)generate plain text
1 <fw>- 15 </fw>
2 véritable d'une hiérarchie fondée sur la rareté. – Je
3 m'amuse parfois d'une image <hi>physique</hi> de nos cœurs,
4 qui sont faits intimement d'une énorme injustice et
5 d'une petite justice combinées. J'imagine qu'il y a
6 dans chacun de nous un atome important entre nos
7 atomes, et constitué par deux <hi>grains d'énergie</hi> qui
8 voudraient bien se séparer. Ce sont des énergies con-
9 tradictoires mais indivisibles. La nature les a jointes
10 pour toujours, quoique furieusement ennemies. L'une
11 est l'éternel mouvement d'un gros <hi>électron positif</hi>, et
12 ce mouvement inépuisable engendre une suite de sons
13 graves où l'oreille intérieure distingue sans nulle peine
14 une profonde phrase monotone: <hi>Il n'y a que moi. Il</hi>
15 <hi>n'y a que moi. I'ln'y a que moi, moi, moi...</hi> Quant au
16 petit électron radicalement négatif, il crie à l'extrême
17 de l'aigu, et perce et reperce de la sorte la plus cruelle
18 le thème égotiste de l'autre : <hi>Oui, mais il ya un tel...</hi>
19 <hi>Oui, mais il y a un, tel... Tel, tel, tel.</hi> Et tel autre !...
20 Car le nom change assez souvent...<p/>
21 <p_/>Bizarre royaume où tentes les belles choses qui s'y
22 produisent sont une amère nourriture pour toutes les
```

Il progetto *Voci della Grande Guerra*

<http://www.vocidellagrandeguerra.it>

Voci della Grande Guerra è un'iniziativa scientifica e culturale promossa dall'Università di Pisa (capofila) in collaborazione con l'Istituto di Linguistica Computazionale "A. Zampolli" del Consiglio Nazionale delle Ricerche (ILC-CNR), l'Università di Siena e l'Accademia della Crusca.

Finanziata dalla Presidenza del Consiglio dei Ministri nell'ambito dell'avviso pubblico per la Commemorazione del Centenario della Grande Guerra, ha l'obiettivo di preservare e diffondere le memorie della Prima Guerra Mondiale attraverso la costruzione e la pubblicazione online di un corpus digitale di testi rappresentativi delle diverse modalità di sentire e raccontare l'Italia in guerra da parte dei suoi protagonisti.

CoPhi Proof-reader per *Voci della Grande Guerra*

Gentile1919 Page 105: Show Image Status: ★★★ Save

Index
OCR

È il motto di una nuova rivista sorta in questi
È il motto di una nuova rivista sorta in questi

[03]

giorni al fronte, da un piccolo nucleo di ufficiali,
giorni al fronte, da un piccolo nucleo di ufficiali,

[04]

che sa d'interpretare l'animo di molti tra i più va-
che sa d'interpretare l'animo di molti tra i più va-

[05]

TEI

Migliorare l'ergonomia

Cos'è l'eye-tracker?

L'*eye tracker* è uno strumento per tracciare il movimento e i punti di fissazione degli occhi su aree di interesse

Esperimenti con l'eye-tracker

Con Barbara Balbi, Flavia De Simone e Vincenzo Broscritto dell'Università "Suor Orsola Benincasa" stiamo conducendo esperimenti con l'*eye tracker* per studiare le diverse strategie di correzione di soggetti non addestrati e di soggetti esperti

Conclusione

- La scelta delle tecniche e degli strumenti di OCR deve essere valutata caso per caso
- Il pre-processing delle immagini con script creati appositamente allo scopo può aumentare notevolmente l'accuratezza
- Il lavoro di correzione può essere ripartito fra diversi soggetti o fra diversi gruppi di lavoro

Domande aperte

- È un modello esportabile?
- Chi dovrebbe guidare progetti di questo tipo?
- Come coinvolgere i ricercatori?
- Come coinvolgere i volontari?
- Come coinvolgere gli insegnanti?
- Come motivare gli studenti?

Grazie per l'attenzione

Bibliografia



S. Feng, R. Manmatha: A Hierarchical, HMM-based Automatic Evaluation of OCR Accuracy for a Digital Library of Books. JCDL 2006, 109–118 (2006)



W.B. Lund, E.K. Ringger: Improving Optical Character Recognition through Efficient Multiple System Alignment, JCDL (2009)



M. Reynaert: Non-interactive OCR Post-correction for Giga-Scale Digitization Projects. A. Gelbukh (ed.): CICLing 2008, LNCS 4919, 617–630 (2008)



M. Reynaert: All, and only, the Errors: more Complete and Consistent Spelling and OCR-Error Correction Evaluation. 6th International Conference on Language Resources and Evaluation 2008, 1867–1872 (2008)



C. Ringlstetter, K. Schulz, S. Mihov, K. Louka: The same is not the same - postcorrection of alphabet confusion errors in mixed-alphabet OCR recognition. 8th International Conference on Document Analysis and Recognition, 1, 406–410 (2005)



M. Spencer, C. Howe: Collating texts using progressive multiple alignment. Computer and the Humanities, 37, 1, 97–109 (2003)



G. Stewart, G. Crane, A. Babeu: A New Generation of Textual Corpora. JCDL 2007, 356–365 (2007)



L. Zhuang, X. Zhu: An OCR Post-processing Approach Based on Multi-knowledge. 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, 346–352 (2005)