(https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality)

ImproveQuality

zdenop edited this page 25 days ago · 27 revisions

Improving the quality of the output 出力品質の向上

There are a variety of reasons you might not get good quality output from Tesseract. It's important to note that unless you're using a very unusual font or a new language retraining Tesseract is unlikely to help.
あなたがTesseractから良い品質の出力を得られないかもしれない様々な理由があります。非常に珍しいフォントやTesseractを再訓練する新しい言語を使用しているのでない限り、役に立ちません。

Image processing 画像処理

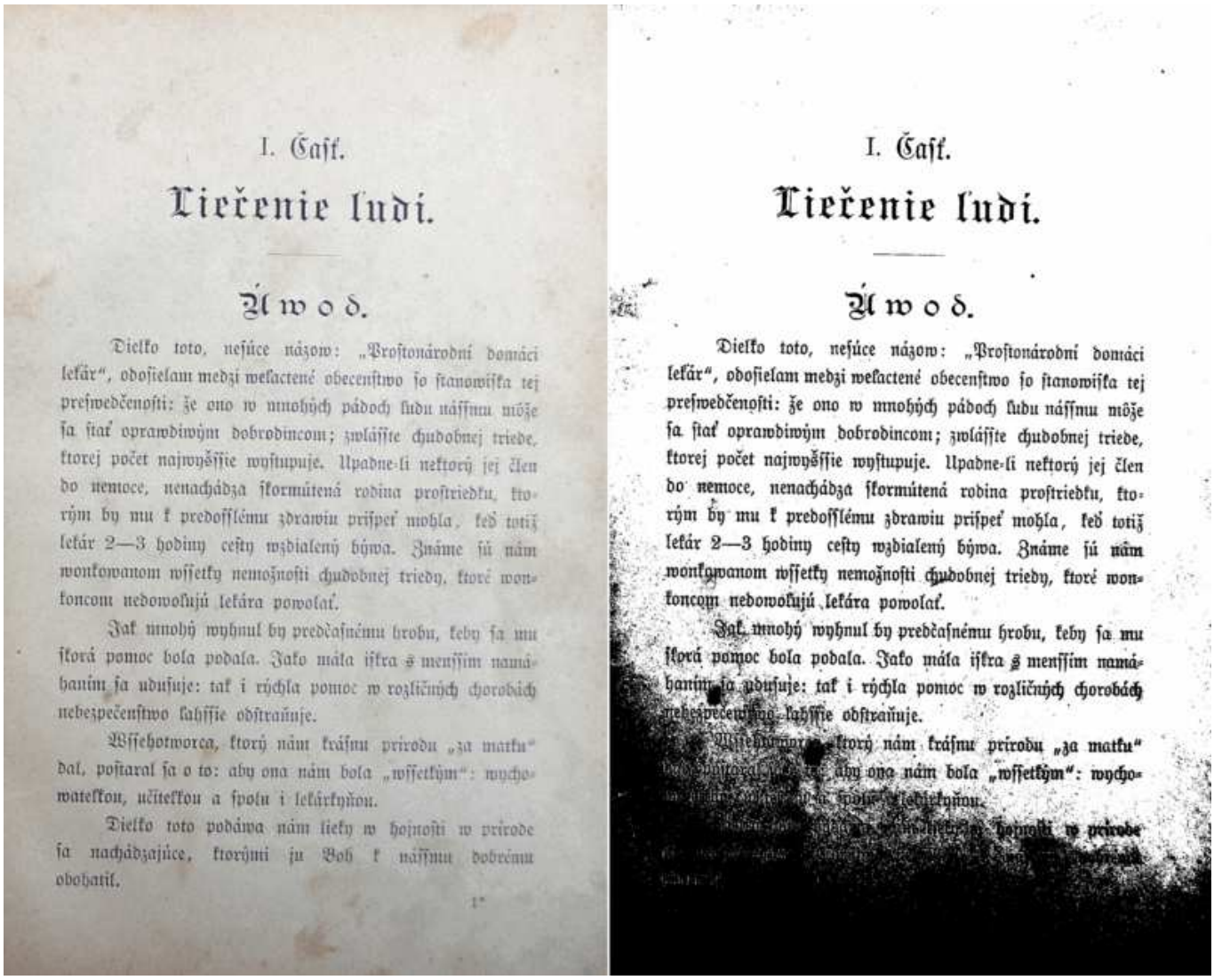
Tesseract does various image processing operations internally (using the Leptonica library) before doing the actual OCR. It generally does a very good job of this, but there will inevitably be cases where it isn't good enough, which can result in a significant reduction in accuracy.
Tesseractは実際のOCRを行う前に内部的に（Leptonicaライブラリを使用して）様々な画像処理操作を行います。それは一般にこれの非常によい仕事をします、しかし、それが十分に良くない場合は必然的にあるでしょう、そしてそれは正確さのかなりの減少をもたらすことができます。

You can see how Tesseract has processed the image by using the configuration variable tessedit_write_images to true when running Tesseract. If the resulting tessinput.tif file looks problematic, try some of these image processing operations before passing the image to Tesseract.
Tesseractの実行時に構成変数tessedit_write_imagesをtrueに設定することで、Tesseractがどのようにイメージを処理したかを確認できます。結果のtessinput.tifファイルに問題があると思われる場合は、Tesseractに画像を渡す前に、これらの画像処理操作のいくつかを試してください。

Rescaling 再スケーリング

Tesseract works best on images which have a DPI of at least 300 dpi, so it may be beneficial to resize images. For more information see the FAQ.
Tesseractは、少なくとも300 dpiのDPIを持つ画像に最適です。そのため、画像のサイズを変更すると効果的です。詳しくはFAQを見てください。

Binarisation 二値化



I. Časť.

Liečenie ľudí.

Úvod.

Dielko toto, nesúce názov: „Prostonárodní domáci lekár“, odosielam medzi welactené obecenstwo so stanowiska tej preswedčenosti: že ono w mnohých pádoch ľudu nášmu môže sa stať oprawdivým dobrodincom; zwlášte chudobnej triede, ktorej počet najwyššie wystupuje. Upadne-li niektorý jej člen do nemoce, nenachádza skormútená rodina prostriedku, ktorým by mu k predošlému zdrawiu prispeť mohla, keď totiž lekár 2—3 hodiny cesty wzdialený býwa. Známe sú nám wonkowanom wšetky nemožnosti chudobnej triedy, ktoré wonkoncom nedowolujú lekára powolať.

Jak mnohý wyhnul by predčasnému hrobu, keby sa mu skorá pomoc bola podala. Jako mála iskra s menšim namáhaním sa udušuje: tak i rýchla pomoc w rozličných chorobách nebezpečenstwo ľahšie odstraňuje.

Wšehotworca, ktorý nám krásnu prirodu „za matku“ dal, postaral sa o to: aby ona nám bola „wšetkým“: wychowateľkou, učiteľkou a spolu i lekárkyňou.

Dielko toto podáwa nám lieky w hojnosti w prirode sa nachádzajúce, ktorými ju Boh k nášmu dobrému obohatil.

1*

I. Časť.

Liečenie ľudí.

Úvod.

Dielko toto, nesúce názov: „Prostonárodní domáci lekár“, odosielam medzi welactené obecenstwo so stanowiska tej preswedčenosti: že ono w mnohých pádoch ľudu nášmu môže sa stať oprawdivým dobrodincom; zwlášte chudobnej triede, ktorej počet najwyššie wystupuje. Upadne-li niektorý jej člen do nemoce, nenachádza skormútená rodina prostriedku, ktorým by mu k predošlému zdrawiu prispeť mohla, keď totiž lekár 2—3 hodiny cesty wzdialený býwa. Známe sú nám wonkowanom wšetky nemožnosti chudobnej triedy, ktoré wonkoncom nedowolujú lekára powolať.

Jak mnohý wyhnul by predčasnému hrobu, keby sa mu skorá pomoc bola podala. Jako mála iskra s menšim namáhaním sa udušuje: tak i rýchla pomoc w rozličných chorobách nebezpečenstwo ľahšie odstraňuje.

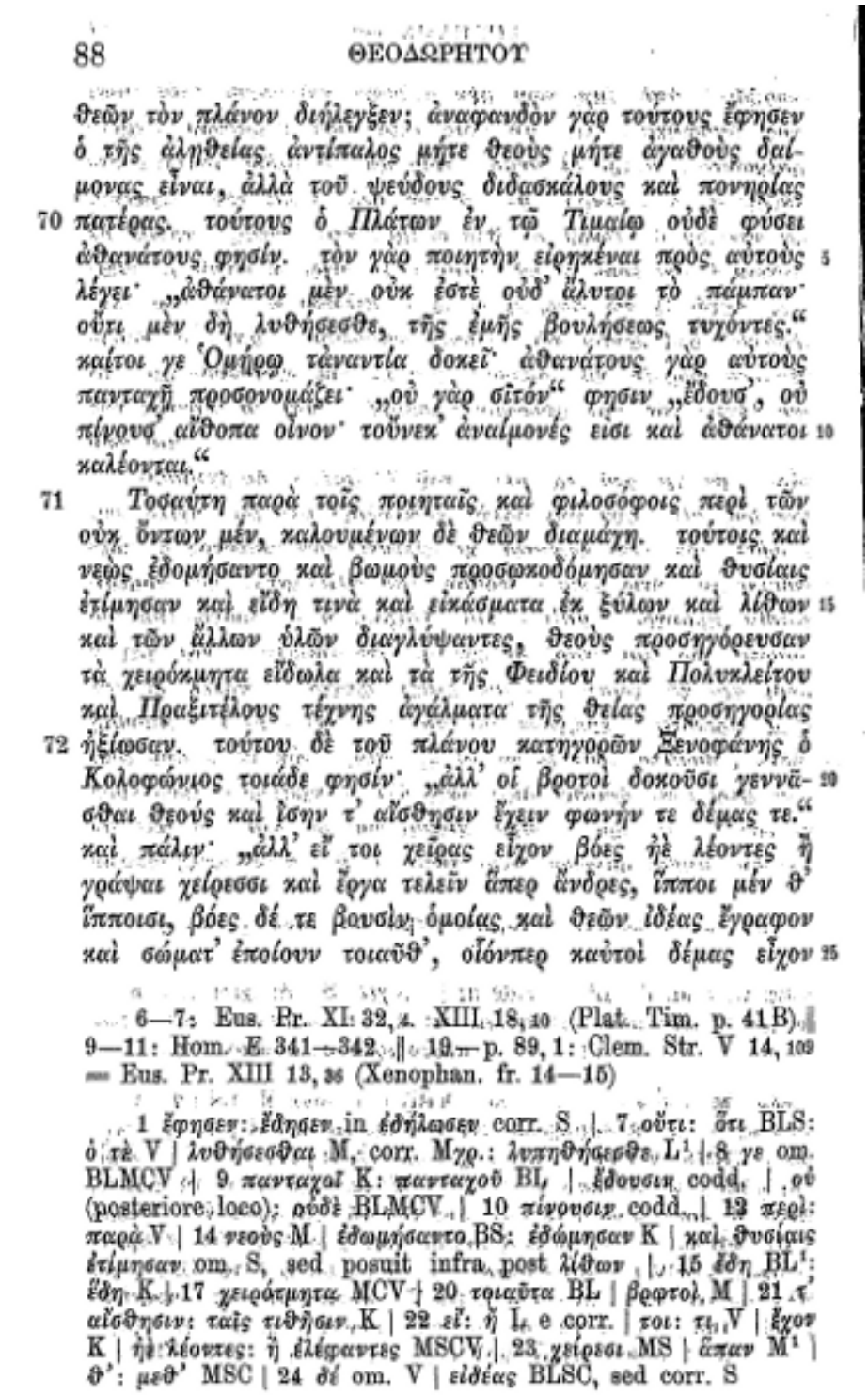
Wšehotworca, ktorý nám krásnu prirodu „za matku“ dal, postaral sa o to: aby ona nám bola „wšetkým“: wychowateľkou, učiteľkou a spolu i lekárkyňou.

Dielko toto podáwa nám lieky w hojnosti w prirode

This is converting an image to black and white. Tesseract does this internally (Otsu algorithm), but the result can be suboptimal, particularly if the page background is of uneven darkness.
これは画像を白黒に変換しています。 Tesseractはこれを内部的に行います（Otsuアルゴリズム）が、特にページの背景が不均一な暗さの場合、結果は最適とは言えません。

If you are not able to fix by better input image, you can try different algorithm. See ImageJ Auto Threshold (java) or OpenCV Image Thresholding (python) or scikit-image Thresholding documentation (python).
あなたがより良い入力画像で修正することができない場合は、別のアルゴリズムを試すことができます。ImageJ Auto Threshold（java）またはOpenCV Image Thresholding（python）またはscikit-image Thresholdingのドキュメント（python）を参照してください。

Noise Removal ノイズ除去



88 ΘΕΟΔΩΡΗΤΟΥ

θεῶν τὸν πλάνον διήλεγξεν; ἀναφανδὸν γὰρ τούτους ἔφησεν
ὁ τῆς ἀληθείας ἀντίπαλος μήτε θεοὺς μήτε ἀγαθοὺς δαί-
μονας εἶναι, ἀλλὰ τοῦ ψεύδους διδασκάλους καὶ πονηρίας
70 πατέρας. τούτους ὁ Πλάτων ἐν τῷ Τιμαίῳ οὐδὲ φύσει
ἀθανάτους φησίν. τὸν γὰρ ποιητὴν εἰρηκέναι πρὸς αὐτοὺς
λέγει· „ἀθάνατοι μὲν οὐκ ἐστὲ οὐδ' ἄλυτοι τὸ πάμπαν·
οὔτι μὲν δὴ λυθήσεσθε, τῆς ἐμῆς βουλήσεως τυχόντες.“
καίτοι γε Ὁμήρῳ τἀναντία δοκεῖ· ἀθανάτους γὰρ αὐτοὺς
πανταχῇ προσονομάζει· „οὐ γὰρ σῖτόν“ φησιν „ἔδουσ', οὐ
πίνουσ' αἴθοπα οἶνον· τοὔνεκ' ἀναιμονές εἰσι καὶ ἀθάνατοι
καλέονται.“

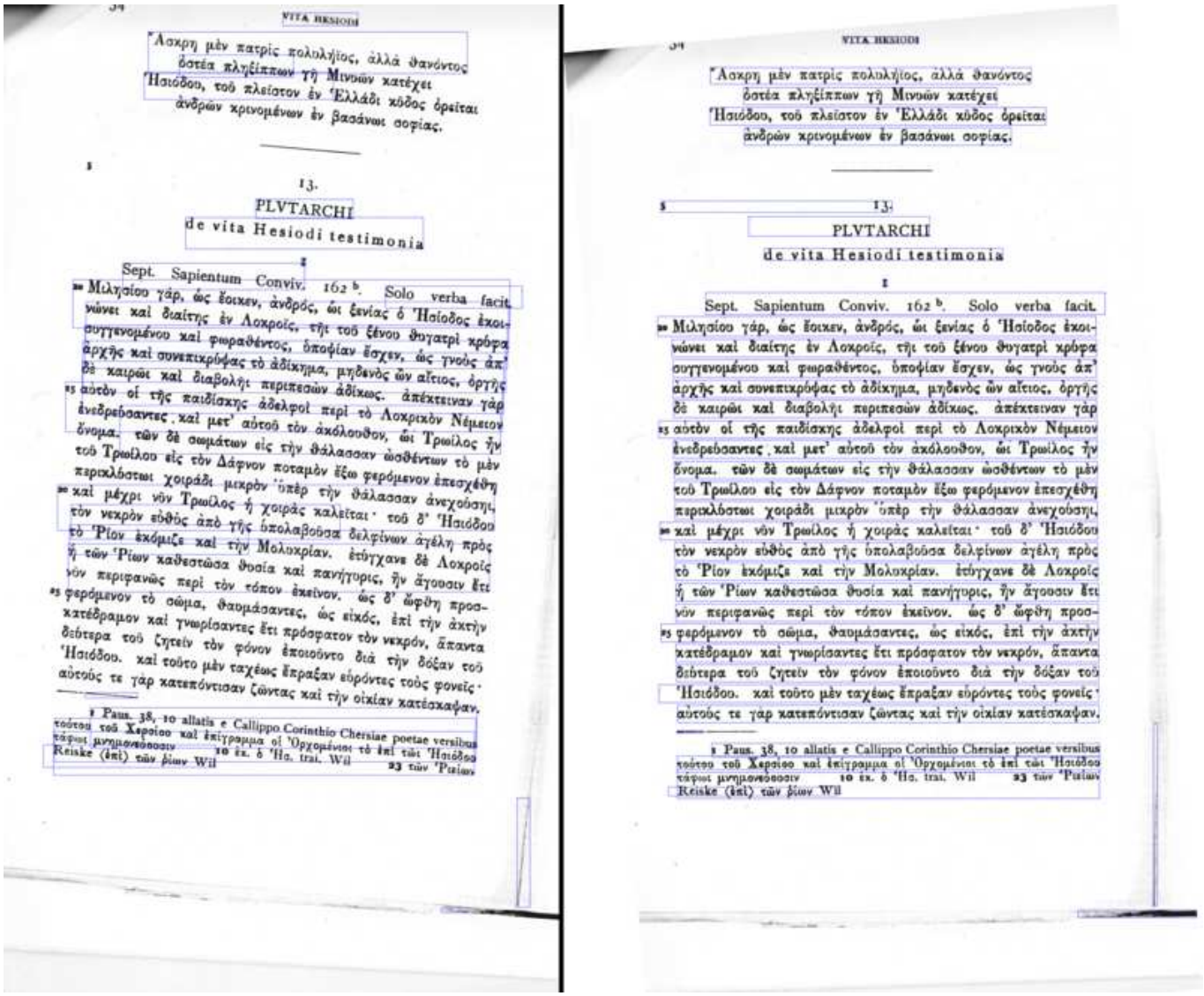
71 Τοσαύτη παρὰ τοῖς ποιηταῖς καὶ φιλοσόφοις περὶ τῶν
οὐκ ὄντων μέν, καλουμένων δὲ θεῶν διαμάχη. τούτοις καὶ
νεὼς ἐδομήσαντο καὶ βωμοὺς προσωκοδόμησαν καὶ θυσίαις
ἐτίμησαν καὶ εἴδη τινὰ καὶ εἰκάσματα ἐκ ξύλων καὶ λίθων
καὶ τῶν ἄλλων ὑλῶν διαγλύψαντες, θεοὺς προσηγόρευσαν
τὰ χειρόκμητα εἴδωλα καὶ τὰ τῆς Φειδίου καὶ Πολυκλείτου
καὶ Πραξιτέλους τέχνης ἀγάλματα τῆς θείας προσηγορίας
72 ἠξίωσαν. τούτου δὲ τοῦ πλάνου κατηγορῶν Ξενοφάνης ὁ
Κολοφώνιος τοιάδε φησίν· „ἀλλ' οἱ βροτοὶ δοκοῦσι γεννᾶ-
σθαι θεούς καὶ ἶσην τ' αἴσθησιν ἔχειν φωνήν τε δέμας τε.“
καὶ πάλιν· „ἀλλ' εἴ τοι χεῖρας εἶχον βόες ἠὲ λέοντες ἢ
γράψαι χείρεσσι καὶ ἔργα τελεῖν ἅπερ ἄνδρες, ἵπποι μέν θ'
ἵπποισι, βόες δέ τε βουσὶν ὁμοίας καὶ θεῶν ἰδέας ἔγραφον
καὶ σώματ' ἐποίουν τοιαῦθ', οἷόνπερ καὐτοὶ δέμας εἶχον

6—7: Eus. Pr. XI 32, 4. XIII 18, 10 (Plat. Tim. p. 41 B). 9—11: Hom. E 341—342. 12—p. 89, 1: Clem. Str. V 14, 109 = Eus. Pr. XIII 13, 36 (Xenophan. fr. 14—15)

1 ἔφησεν: ἔδησεν in ἐδήλωσεν corr. S | 7 οὔτι: ὅτι BLS: ὁ τὲ V | λυθήσεσθαι M, corr. Mχρ.: λυπηθήσεσθε L¹ | 8 γε om. BLMCV | 9 πανταχαῖ K: πανταχοῦ Bl, | ἔδουσιν codd. | οὐ (posteriore loco): οὐδὲ BLMCV | 10 πίνουσιν codd. | 12 περὶ: παρὰ V | 14 νεοὺς M | ἐδωμήσαντο BS: ἐδώμησαν K | καὶ θυσίαις ἐτίμησαν om. S, sed posuit infra post λίθων | 15 εἴδη BL¹: ἔδη K | 17 χειράτμητα MCV | 20 τοιαῦτα BL | βρφτοὶ M | 21 τ' αἴσθησιν: ταῖς τιθῆσιν K | 22 εἴ: ἤ L, e corr. | τοι: τι V | ἔχον K | ἠὲ λέοντες: ἢ ἐλέφαντες MSCV | 23 χείρεσι MS | ἅπαν M¹ | θ': μεθ' MSC | 24 δέ om. V | εἰδέας BLSC, sed corr. S

Noise is random variation of brightness or colour in an image, that can make the text of the image more difficult to read. Certain types of noise cannot be removed by Tesseract in the binarisation step, which can cause accuracy rates to drop.
ノイズは、画像内の明るさや色のランダムな変動で、画像のテキストを読みにくくすることがあります。2値化ステップでTesseractで特定の種類のノイズを除去することはできません。これにより、正解率が低下する可能性があります。

Rotation / Deskewing 回転 / 傾き補正



VITA HESIODI

Ἄσκρη μὲν πατρὶς πολυλήϊος, ἀλλὰ θανόντος
δοτέα πλῆξίππων γῆ Μινυῶν κατέχει
Ἡσιόδου, τοῦ πλείστον ἐν Ἑλλάδι κῦδος ὀρεῖται
ἀνδρῶν κρινομένων ἐν βασάνωι σοφίας.

13.

PLVTARCHI
de vita Hesiodi testimonia

Sept. Sapientum Conviv. 162 b. Solo verba facit.

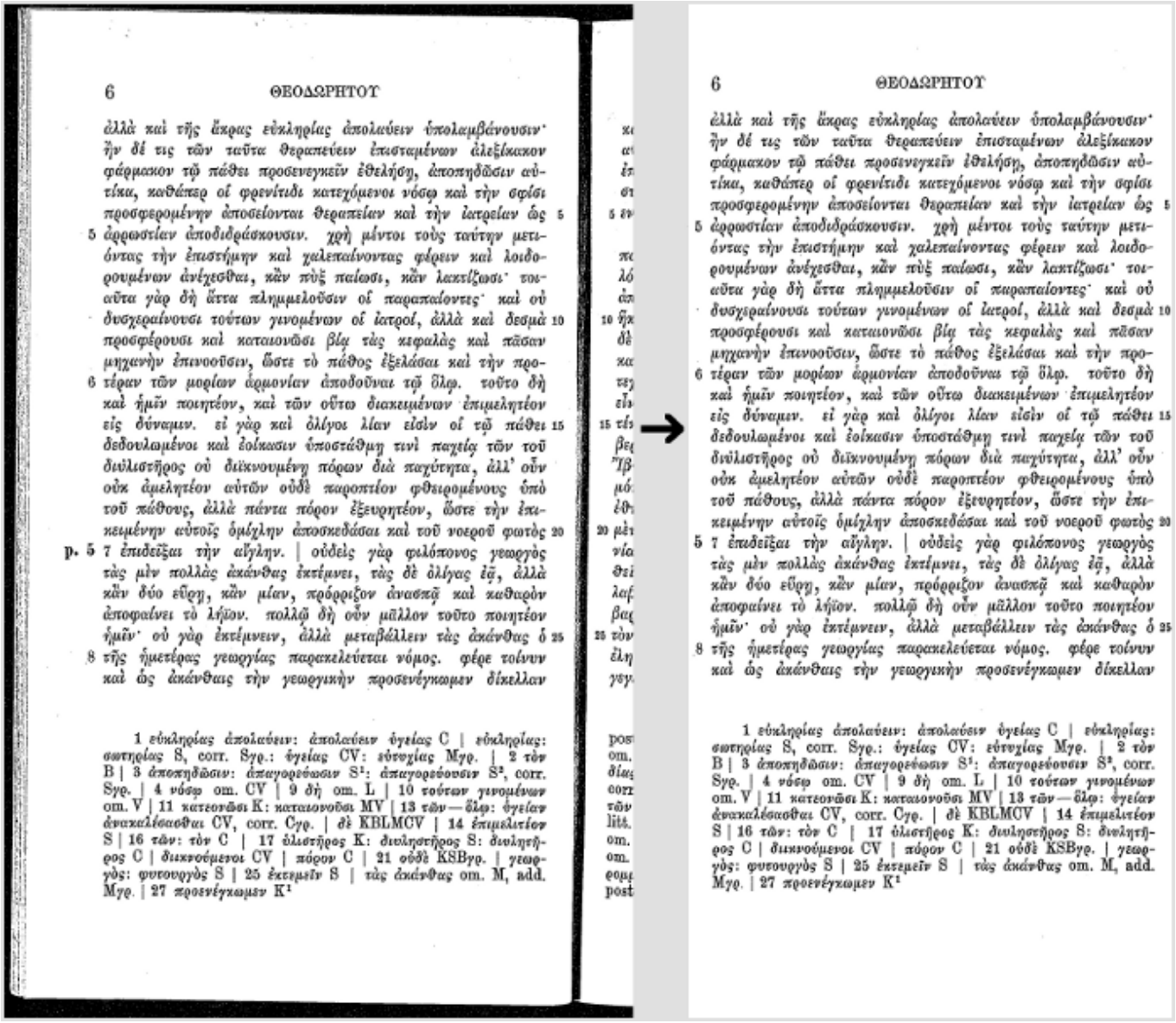
Μιλησίου γάρ, ὡς ἔοικεν, ἀνδρός, ὠι ξενίας ὁ Ἡσίοδος ἐκοινώνει καὶ διαίτης ἐν Λοκροῖς, τῆι τοῦ ξένου θυγατρὶ κρύφα συγγενομένου καὶ φωραθέντος, ὑποφίαν ἔσχεν, ὡς γνοὺς ἀπ' ἀρχῆς καὶ συνεπικρύψας τὸ ἀδίκημα, μηδενὸς ὢν αἴτιος, ὀργῆς δὲ καιρῶι καὶ διαβολῆι περιπεσὼν ἀδίκως. ἀπέκτειναν γάρ αὐτὸν οἱ τῆς παιδίσκης ἀδελφοὶ περὶ τὸ Λοκρικὸν Νέμειον ἐνεδρεύσαντες, καὶ μετ' αὐτοῦ τὸν ἀκόλουθον, ὠι 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 ex. ὁ Ἡσ. trai. Wil 23 τῶν Ῥιείων Reiske (ἐπὶ) τῶν βίων Wil

A skewed image is when a page has been scanned when not straight. The quality of Tesseract's line segmentation reduces significantly if a page is too skewed, which severely impacts the quality of the OCR. To address this rotating the page image so that the text lines are horizontal.
画像が歪んでいると、ページがまっすぐでないときにスキャンされたときです。 ページが歪んでいると、Tesseractのラインセグメンテーションの品質が大幅に低下します。これはOCRの品質に深刻な影響を与えます。 これに対処するには、テキスト行が水平になるようにページ画像を回転させます。

Borders 境界

Scanning border Removal スキャニングボーダー除去



Scanned pages often have dark borders around them. These can be erroneously picked up as extra characters, especially if they vary in shape and gradation.
スキャンされたページは、その周囲に暗い縁があることがよくあります。特に形やグラデーションが異なる場合、これらは誤って余分な文字として認識される可能性があります。

Missing borders 余白が無い

If you OCR just text area without any border, tesseract could have problems with it. See for some details in [tesseract user forum#427](#) . You can easy add small border (e.g. 10 pt) with [ImageMagick®](#):
余白無しのテキスト領域だけをOCRした場合、tesseractに問題が生じる可能性があります。
tesseractユーザー・フォーラム＃427の詳細についてはを参照してください。あなたは簡単にImageMagick®で小さいボーダー（例えば10 pt）を加えることができます：

```
convert  427-1.jpg  -bordercolor White -border 10x10 427-1b.jpg
```

Transparency / Alpha channel 透明度 / アルファチャンネル

Some image formats (e.g. png) can have alpha-channel for providing transparency feature.
一部の画像フォーマット（例：png）は、透明機能を提供するためのアルファチャンネルを持つことができます。

Tesseract 3.0x expects that users remove alpha channel from image before using image in tesseract. This can done e.g. with ImageMagick command:
Tesseract 3.0xはユーザーがtesseractで画像を使用する前に画像からアルファチャンネルを削除することを期待しています。これはできます。 ImageMagickコマンドで：

```
convert input.png -alpha off output.png
```

Tesseract 4.00 removes alpha channel with leptonica function pixRemoveAlpha(): it removes alpha component by blending with white background. In some case (e.g. OCR of movie subtitles) this can lead to problems, so users would need to remove alpha channel (or pre-process image by inverting image colors) by themself.
Tesseract 4.00はレプトニカ関数pixRemoveAlpha()でアルファチャンネルを削除します。白い背景とブレンドすることでアルファ成分を削除します。場合によっては（例えば、映画の字幕のOCR）、問題が生じる可能性があるので、ユーザは自分でアルファチャンネルを削除する（または画像の色を反転させることによって画像を前処理する）必要があるだろう。

Tools / Libraries ツール / ライブラリ

- Leptonica
- OpenCV
- Scan Tailor
- ImageMagick
- unpaper
- ImageJ
- Gimp

Examples 例

If you need an example how to improve image quality programmatically, have a look at this examples:
プログラムで画質を向上させる方法の例が必要な場合は、次の例を見てください。

- OpenCV - Rotation (Deskewing) - c++ example
  OpenCV - 回転（傾き補正） - c++の例

- Fred's ImageMagick TEXTCLEANER - bash script for processing a scanned document of text to clean the text background.
  FredのImageMagick TEXTCLEANER - テキストのスキャンされた文書を処理してテキストの背景をきれいにするためのbashスクリプト。

- rotation_spacing.py - python script for automatic detection of rotation and line spacing of an image of text
  rotation_spacing.py - テキストの画像の回転と行間隔を自動検出するためのpythonスクリプト

- crop_morphology.py - Finding blocks of text in an image using Python, OpenCV and numpy
  crop_morphology.py - Python、OpenCV、およびnumpyを使用して画像内のテキストブロックを見つける

- Credit card OCR with OpenCV and Python OpenCVとPythonのクレジットカードOCR

- noteshrink - python example how to clean up scans. Details in blog Compressing and enhancing hand-written notes.
  noteshrink - スキャンをクリーンアップする方法のpythonの例。ブログでの詳細手書きのメモの圧縮と強化。

- uproject text - python example how to recover perspective of image. Details in blog Unprojecting text with ellipses.
  uproject text - 画像のパースペクティブを回復する方法のpythonの例。ブログの詳細省略記号付きのテキストの投影を解除する。

- page_dewarp - python example for Text page dewarping using a "cubic sheet" model. Details in blog Page dewarping.
  page_dewarp - "cubic sheet"モデルを使ったテキストページの歪み補正のpythonの例。詳細はブログのページdewarpingで。

Page segmentation method ページ分割方法

By default Tesseract expects a page of text when it segments an image. If you're just seeking to OCR a small region try a different segmentation mode, using the --psm argument. Note that adding a white border to text which is too tightly cropped may also help, see issue 398.
デフォルトでは、Tesseractは画像を分割するときに1ページのテキストを期待します。 OCRで小さな領域を探しているだけの場合は、 - psm引数を使用して別のセグメンテーションモードを試してください。きちんとトリミングされているテキストに白いボーダーを追加することも助けになるかもしれないことに注意してください、問題398を見てください。

To see a complete list of supported page segmentation modes, use tesseract -h. Here's the list as of 3.21:
サポートされているページセグメンテーションモードの完全なリストを見るには、tesseract -hを使用してください。これは3.21の時点でのリストです:

```
 0    Orientation and script detection (OSD) only.
      方向と文字検出(OSD)のみ。

 1    Automatic page segmentation with OSD.
      OSDによる自動ページセグメンテーション。

 2    Automatic page segmentation, but no OSD, or OCR.
      自動ページ分割。ただし、OSD、またはOCRはありません。

 3    Fully automatic page segmentation, but no OSD. (Default)
      全自動ページセグメンテーション、OSDなし。(デフォルト)

 4    Assume a single column of text of variable sizes.
      可変サイズの1列のテキストを想定します。

 5    Assume a single uniform block of vertically aligned text.
      垂直方向に配置されたテキストの単一の均一なブロックを仮定します。

 6    Assume a single uniform block of text.
      一様なテキストブロックを仮定する。

 7    Treat the image as a single text line.
      画像を1行のテキストとして扱います。

 8    Treat the image as a single word.
      画像を一語として扱います。

 9    Treat the image as a single word in a circle.
      画像を円の中の1つの単語として扱います。

10    Treat the image as a single character.
      画像を1文字として扱います。

11    Sparse text. Find as much text as possible in no particular order.
      スパーステキスト。特定の順序でできるだけ多くのテキストを見つけます。

12    Sparse text with OSD.
      OSD付きのまばらなテキスト。

13    Raw line. Treat the image as a single text line,
       bypassing hacks that are Tesseract-specific.
      生の行画像を1行のテキストとして扱い、Tesseract固有のハッキングを回避する。
```

Dictionaries, word lists, and patterns
辞書、単語リスト、およびパターン

By default Tesseract is optimized to recognize sentences of words. If you're trying to recognize something else, like receipts, price lists, or codes, there are a few things you can do to improve the accuracy of your results, as well as double-checking that the appropriate segmentation method is selected.
デフォルトでは、Tesseractは単語の文章を認識するように最適化されています。領収書、価格表、コードなど、他のものを認識しようとしている場合は、結果の精度を向上させるためにできることがいくつかあります。また、適切なセグメンテーション方法が選択されていることを再確認します。

Disabling the dictionaries Tesseract uses should increase recognition if most of your text isn't dictionary words. They can be disabled by setting the both of the configuration variables load_system_dawg and load_freq_dawg to false.
テキストの大部分が辞書の単語ではない場合、Tesseractが使用する辞書を無効にすると、認識が向上します。設定変数load_system_dawgとload_freq_dawgの両方をfalseに設定することでそれらを無効にすることができます。

It is also possible to add words to the word list Tesseract uses to help recognition, or to add common character patterns, which can further help to improve accuracy if you have a good idea of the sort of input you expect. This is explained in more detail in the Tesseract manual.
認識を容易にするためにTesseractが使用する単語リストに単語を追加したり、一般的な文字パターンを追加したりすることもできます。これはTesseractマニュアルでより詳しく説明されています。

If you know you will only encounter a subset of the characters available in the language, such as only digits, you can use the tessedit_char_whitelist configuration variable. See the FAQ for an example.
数字だけのように、言語で利用可能な文字のサブセットのみに出会うことがわかっている場合は、tessedit_char_whitelist構成変数を使用できます。例としてFAQを見てください。

Still having problems? まだ問題がありますか？

If you've tried the above and are still getting low accuracy results, ask on the forum for help, ideally posting an example image.
上記を試してもまだ精度が低い結果が得られる場合は、フォーラムで質問してください。理想的にはサンプル画像を投稿してください。

Expect to add LCD font library。! LCDフォントライブラリを追加する予定です。