The Fourth SEEDI Conference
Digitization of cultural and scientific heritage
Belgrade, Serbia, June 12-15, 2008

A software tool for searching in binary text images

Nikolay Kirov Kirov

Computer Science Department, NBU and
Institute of Mathematics and Informatics, BAS
Sofia, Bulgaria

Abstract: In this paper we present a software tool for searching word images in scanned text documents. We consider that the document pages are represented as files in tif, jpg, gif, png, bmp and other graphic file formats. Our experiments prove the efficiency of the proposed approach and show that such type of search could be successful. Examples of using various languages are presented. The software is user oriented and can be applied to any collection of scanned documents.

This research has been partially supported by a Marie Curie Fellowship of the European Community program "Knowledge Transfer for Digitalization of Cultural and Scientific Heritage in Bulgaria" under contract number MTKD-CT-2004-509754.

Introduction

Optical character recognition (OCR) is the usual way of conducting text retrieval from scanned document images.

It converts text images into a text file, recognizing every symbol and mapping it to a number, which is called code.

The most often used codes are ASCII (one byte code) or UTF-8 (two bytes code).

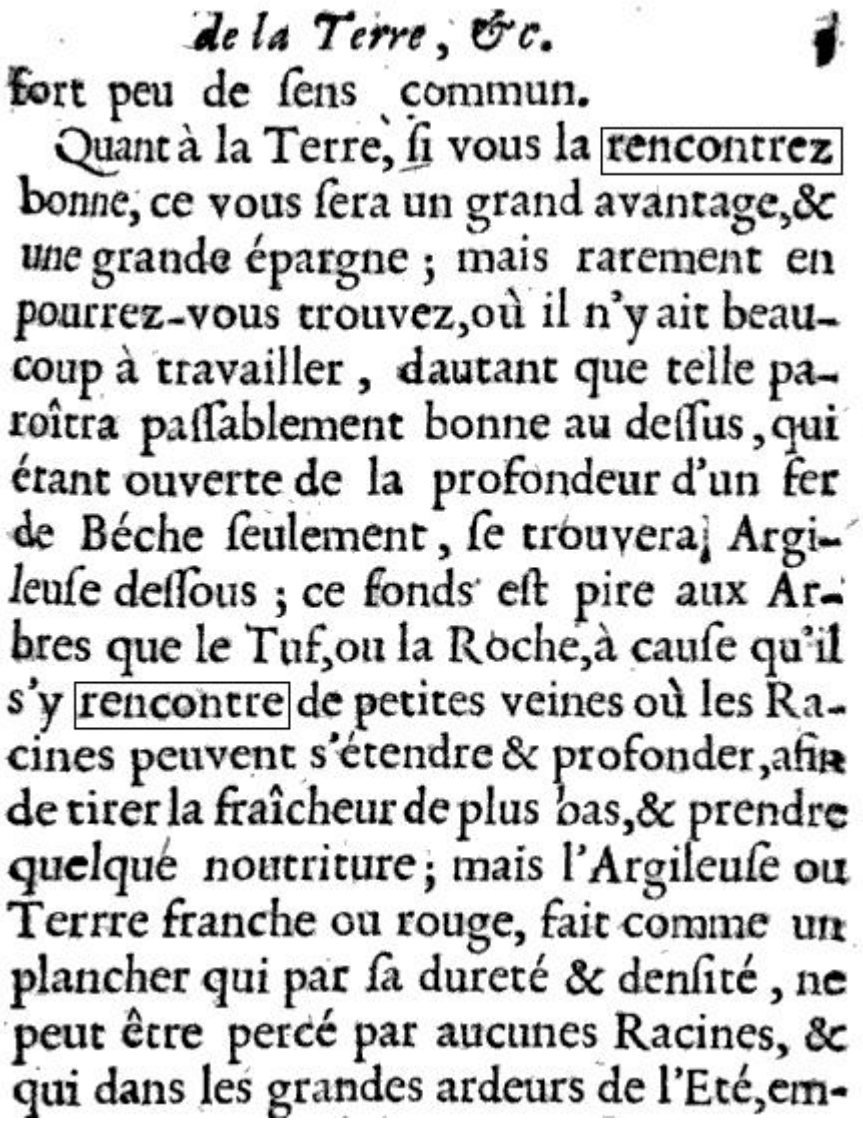
This technique is well developed and has high accuracy.

Searching words in a text file is relatively easy task.

But sometimes OCR is a very difficult process requiring dictionaries in the corresponding languages. Often human efforts are needed to correct OCR errors which is quite tedious work. There are some obstacles to successful OCR:

- The quality of page images.
- Language dependency (alphabet and coding, unknown language):
  - dictionaries;
  - old grammar, obsolete words and phrases and idioms;
  - old letters, outside of the coding tables;
  - multi-lingual documents;
- Errors in automatic OCR, human intervention needed.

Nicolas de Bonnefons, Ch. de Sergy, (1692), University of Gent, Digitized by Google (2007)



de la Terre, &c.

fort peu de ſens commun.

Quant à la Terre, ſi vous la rencontrez bonne, ce vous ſera un grand avantage, & une grande épargne ; mais rarement en pourrez-vous trouvez, où il n'y ait beaucoup à travailler, dautant que telle paroîtra paſſablement bonne au deſſus, qui étant ouverte de la profondeur d'un fer de Béche ſeulement, ſe trouvera Argileuſe deſſous ; ce fonds eſt pire aux Arbres que le Tuf, ou la Roche, à cauſe qu'il s'y rencontre de petites veines où les Racines peuvent s'étendre & profonder, afin de tirer la fraîcheur de plus bas, & prendre quelque nourriture ; mais l'Argileuſe ou Terrre franche ou rouge, fait comme un plancher qui par ſa dureté & denſité, ne peut être percé par aucunes Racines, & qui dans les grandes ardeurs de l'Eté, em-

n»il)uanrΓ¤, la l'erre,^! vouz la rencomrex
t,onne,ce vouz lera un ^ranΓ¤ avanra^e,sc
u/B»e ^ranΓ¤B« epar^ne ; maiz raremenr en
pourrex-vouz trouvex,on il n'v air tieau-
couri a travailler , claur^nc ^ue teile ria-
rorrra rrallΓ¤lllernenr cionne au clelluz,c^
am ouverre cle la rirokoncleur Γ¤'un iec
Γ¤e Lecrre ieulemenr , re rrouveral ^r?l-
leule cleclouz ; ce fonclz eli riire aux ^r-
Γ¶rez c^ue le l"uf,on la l^t,cke,a caule c^u'il
z'^ reuconrre cle r/erirez veiuez on lez ^.a-
cinez peuvenr z'ccenΓ¤resc rirofoncler,arlB«
cle rirer la ir,uclleurcle^luz baz,sc rirenΓ¤re ^
uelc^ue nourricure; malz l^r^ilcule ou '
I'errre francrre ou rou^e, fair comme un
plancrrer c^ui riar ia Γ¤urere sc clenlrre , ne
peur ecre rierce riar aucunez li.acinez, sc
c^ui cl^nz lez ^r,inclez arcleurz cle l'^,te,emB»

We suggest a different approach: instead of applying two steps – OCR and searching in text documents, we will directly search words in scanned text documents.

We can organize retrieval of words, similar to a given pattern word, searching in the binary text images.

The document pages can be represented as binary images in any graphic file format.

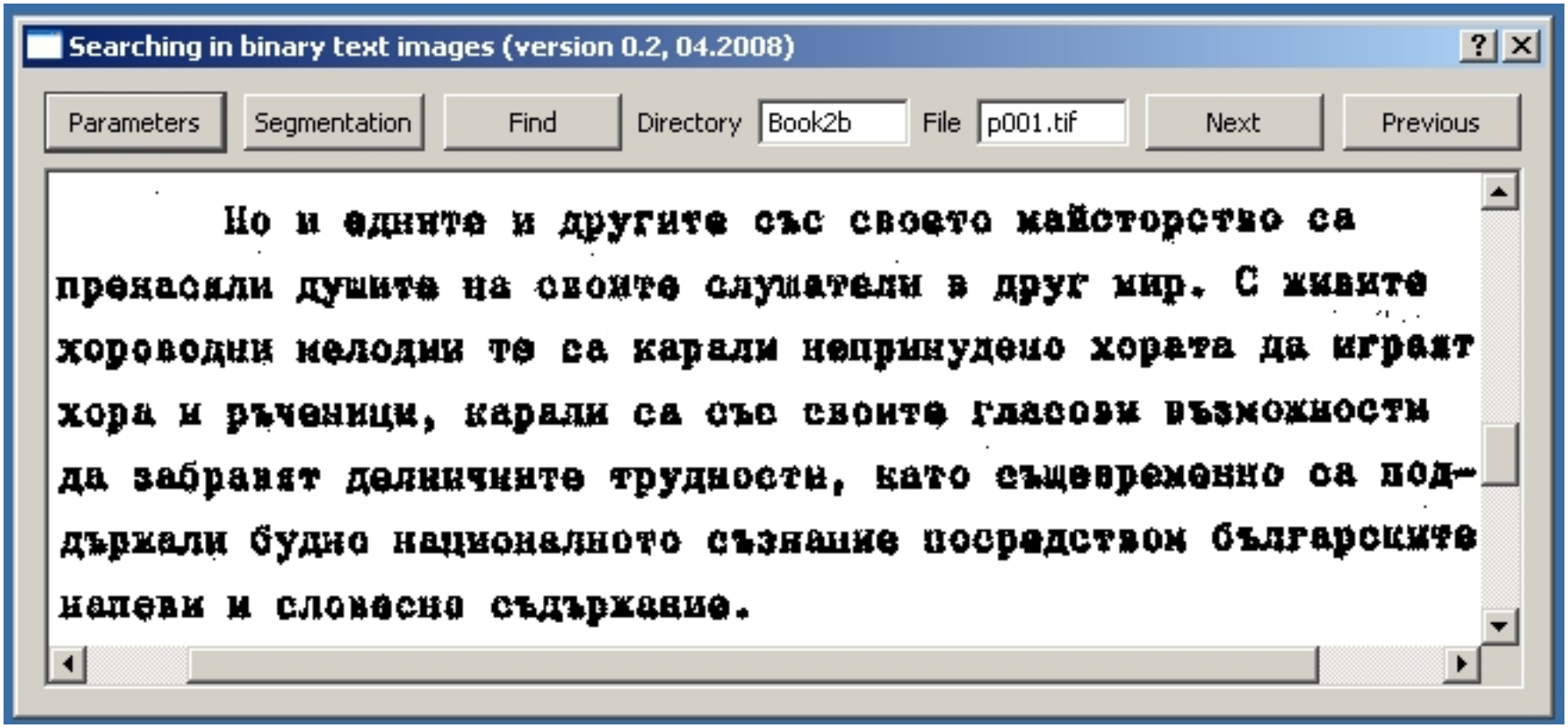
Input data of the software are collection of files representing text document.

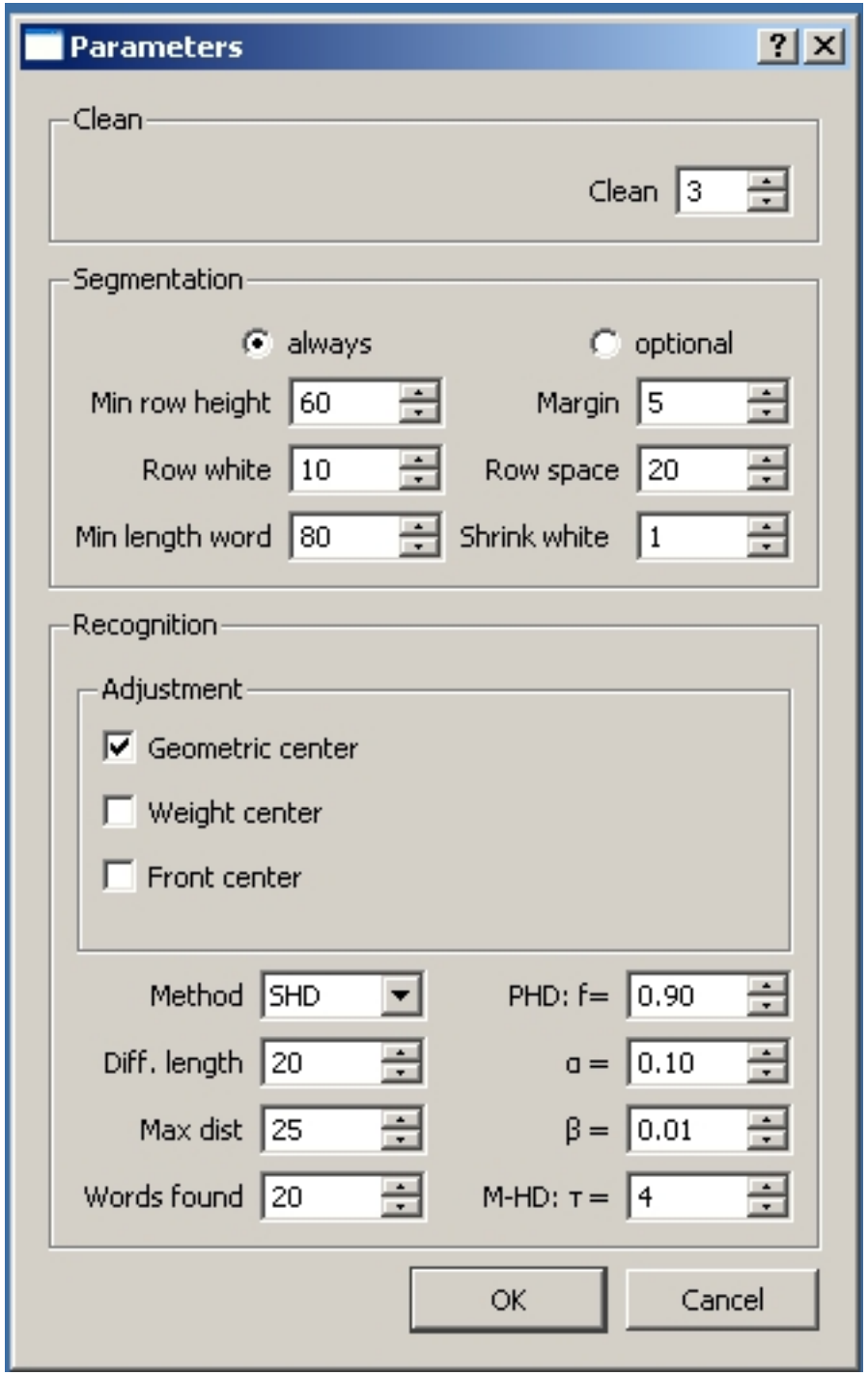
The software

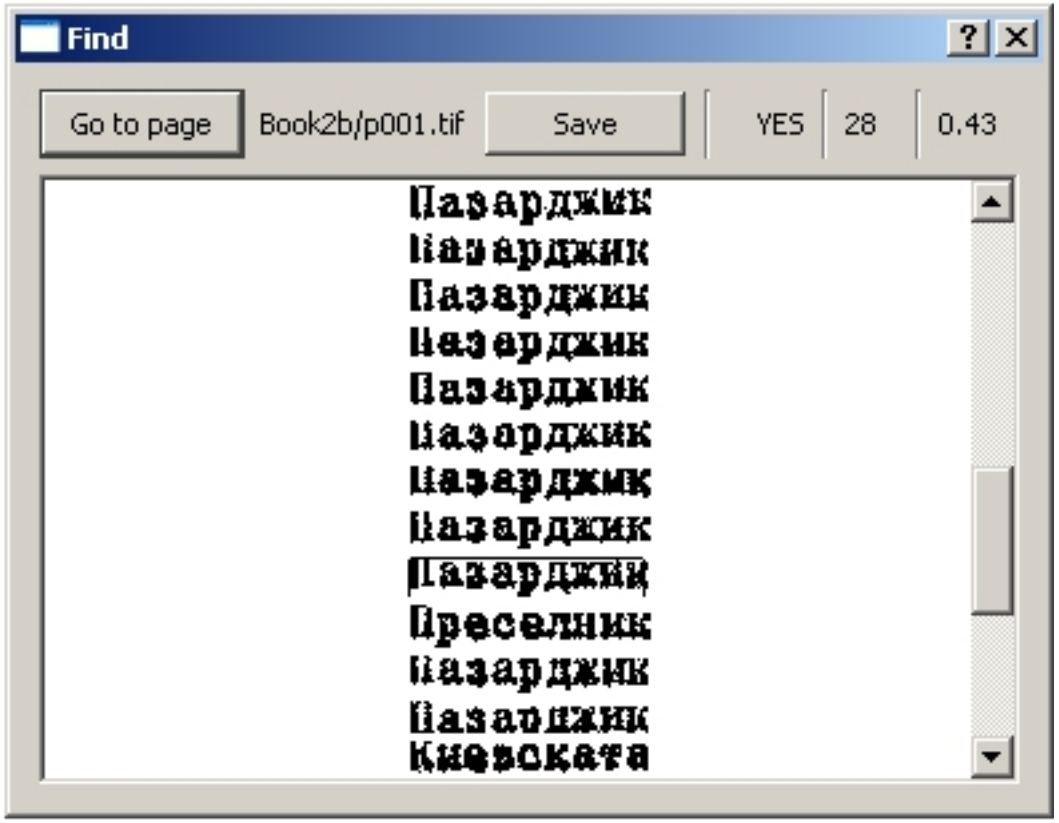
The software system supports three user windows:

- `Main`
- `Parameters`
- `Found`.

`Main` window displays one page of the document – the current page. The current directory containing document image files and the current file name are given on the top of `Main` window. It is possible to go forward and backward through the document pages.




Parameters
Clean
Clean 3
Segmentation
always
optional
Min row height 60
Margin 5
Row white 10
Row space 20
Min length word 80
Shrink white 1
Recognition
Adjustment
Geometric center
Weight center
Front center
Method SHD
PHD: f= 0.90
Diff. length 20
α = 0.10
Max dist 25
β = 0.01
Words found 20
M-HD: τ = 4
OK
Cancel


Find
Go to page
Book2b/p001.tif
Save
YES
28
0.43
Пазарджик
Пазарджик
Пазарджик
Пазарджик
Пазарджик
Пазарджик
Пазарджик
Пазарджик
Пазарджик
Преселник
Пазарджик
Пазарджик
Киевската

Segmentation

We use horizontal projection for line extraction. If the lines are horizontal straight lines or have small slopes, the histogram has near zero values between lines.

To segment the words in a line, we use vertical projection – a histogram obtained by counting the number of black pixels in each vertical scan at a given horizontal position. If the words are well separated, the histogram should have areas of zero values between words.

As a result of the segmentation, every word is associated with a word image – minimal rectangular frame that contains the corresponding word. So we consider any word image as a rectangle, which consists of white and black pixels. The black pixels form a set, which is used in calculating word similarities.

For segmentation step we use a number of parameters, which are important for successful word separation:

- **`Minimal row height`**: This paramter helps us to avoid creating (due to noise) rows with small height;
- **`Margin`**: This paramter allows us to process only a part of the page.
- **`Row white`**: When the value at a point in row histogram is less than the value of this parameter, we suppose that this point belongs to the white space between the words.
- **`Row space`**: Minimum value of the white space between words.
- **`Minimum word length`**: The system does not segment words with length less than the value of this parameter. **`Shrink white`**: This parameter concerns additional step conducted when we have already separated words, and words are framed. At this step we try to shrink the height of rectangles if the points of histograms have values less than this parameter.

Searching

After segmentation of a page, we must choose a pattern word image – this is a word, which we want to find in the document pages.

This step starts when the user push the button `Find` in `Main` window. It activates the process of inspection all pages to measure similarities of segmented words and the pattern word.

We can see a part of the retrieval data in `Find` window. Pushing `GoTo` button, the page containing the marked word downloads in `Main` window.

The program code is written in C++ with help of Qt – a cross-platform application development framework (open soutce from Trolltech).

Experiments

Bulgarian typewritten document (about 1940), 335 pages, tif (2400×3200), 1 BPP

I РАЗДЕЛ

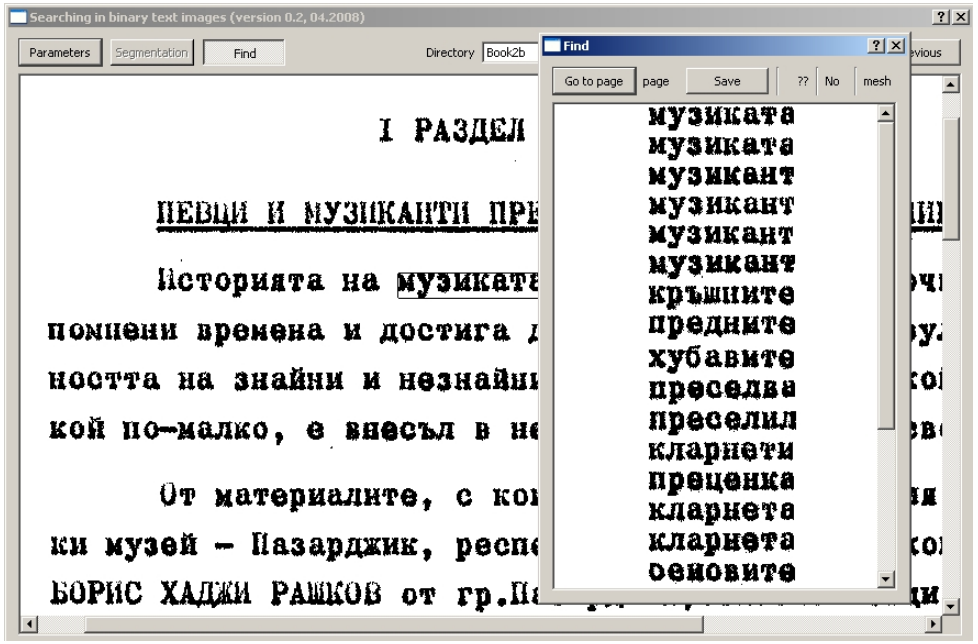
ПЕВЦИ И МУЗИКАНТИ ПРЕДИ И СЛЕД ОСВОБОЖДЕНИЕТО

Историята на музиката в гр.Пазарджик започва от незапомнени времена и достига до наши дни, като резултат от дейността на знайни и незнайни труженици, които, кой повече, кой по-малко, е внесъл в нейната съкровищница своя дял.

От материалите, с които разполага Окръжния исторически музей – Пазарджик, респективно сведенията, които е събрал БОРИС ХАДЖИ РАШКОВ от гр.Пазарджик,относно певци и музиканти преди и след Освобождението се установява, че битовите нужди, свързани с годежи, сватби, занаятчийско-еснафски сбирки, хора, вечеринки и пр. са били задоволявани от музиканти – професионалисти и любители.

Професионалисти били онези музиканти-инструменталисти или певци, като най-често инструменталиста е бил и певец, които са свирили и пеели срещу възнаграждение, а любители-онези, които със своето пеене и свирене са радвали душите и сърцата на хората по сборове, хорища и др., без да получават възнаграждение.

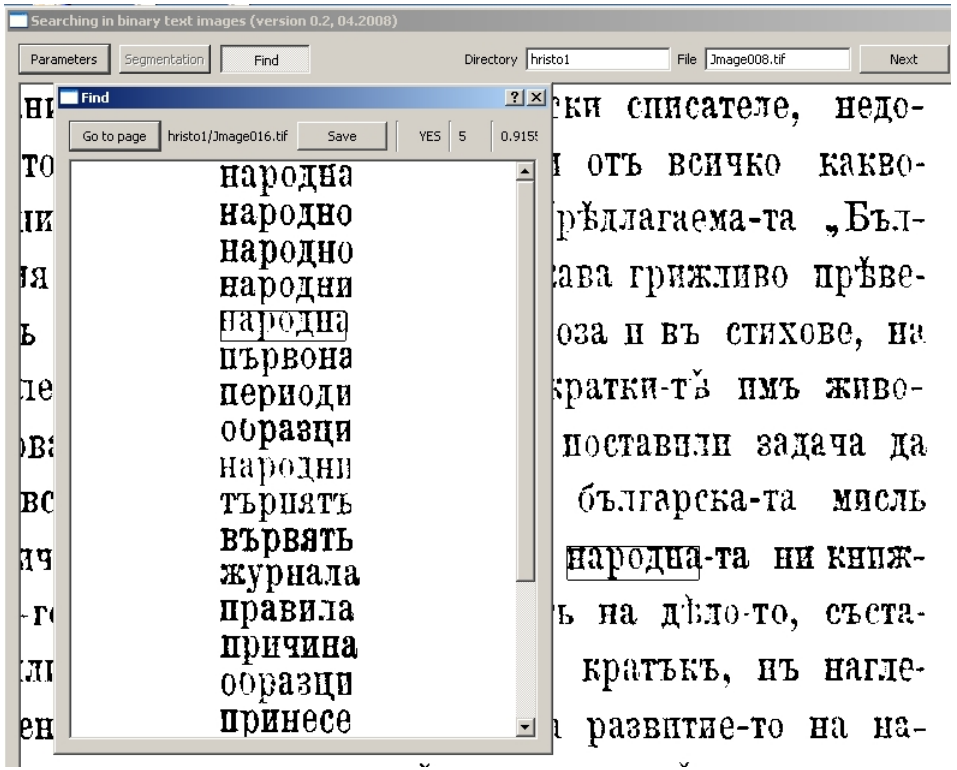
Но и едните и другите със своето майсторство са пренасяли душите на своите слушатели в друг мир. С живите хороводни мелодии те са карали непринудено хората да играят хора и ръченици, карали са със своите гласови възможности да забравят делничните трудности, като същевременно са поддържали будно националното съзнание посредством българските напеви и словесно съдържание.



Bulgarian book, Christomatia (1884), 1000 pages, tif (2300×3800), 8 BPP

Повече-то отъ ранни-тѣ му стихотворения сж любовни пѣсни, по подражание на гръцки-тѣ, и не прѣдставлявать литературна стойность; стихотворения-та му въ „Смѣсна Китка“ при всичко, че повечето сж слаби подражания на руски-тѣ, нъ свидѣтелствовать вече за поетическо-то дарование на г. Славейкова: най-добри-тѣ му стихотворения сж обнародвани-тѣ по-послѣ въ „Читалиште,“ отъ които „Не пѣй ми се,“ „Жестокость-та ми се сломи“ и „Тогасъ понѣ“ джхать съ истински лиризмъ и заслужено привлѣкоха внимание-то на читателе-тѣ. Славейковъ, който е вѣшть въ български езикъ, пръвъ доказа гжвкость-та му въ поезия-та. Като се числи между първи-тѣ борци по черковний въпросъ, той захваща въ сжито-то врѣме почтенно мѣсто въ редъ-тъ на малко-то ни добри литератори.

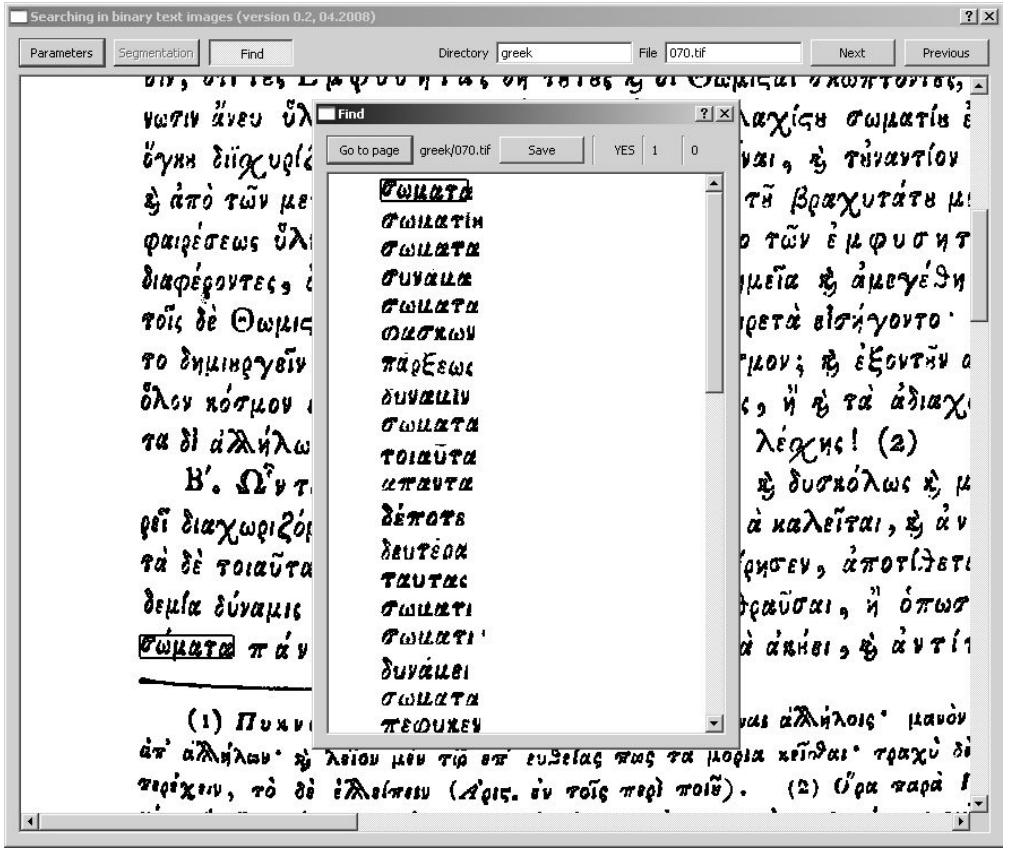
Велико влияние е упражнила възъ пробуждание-то духъ-тъ камъ свобода-та на независимость-та у български народъ доста обширна-та литературна дѣятелность на *Георги Сава Раковски* (род. въ Котель 1818, умр. въ Букурешь 1868 г.). Въ личность-та и въ дѣла-та на Раковски се отрази най-нагледно тогавашно-то състояние на умове-тѣ, нужди-тѣ, стремления-та и идеали-тѣ на народъ-тъ ни. Тако-речи единчькъ дѣець по онова врѣме, той писува, работи всичко. Той искаше да обгърне въ своя-та широка дѣятелность всички-тѣ нужди на народъ-тъ ни, да удовлетвори всички-тѣ национални купнѣяния, да осжществи най-смѣлии-тѣ и въжделени мечти. Той възсъздаде съ фанатический въсторгъ миижло-то и приготви бжджите-то. Бѣше въ сжито-то врѣме поетъ, историкъ, етнографъ, публицистъ, агитаторъ и хайдутинъ. Нито на единъ български дѣятель животъ-тъ не е билъ напълненъ съ толкова неутолима и разнообразна дѣятелность и напьстренъ съ толкова бѣди, приключения и странности. Той се бѣше училъ въ Атина, Парижь, Цариградъ и въ Русия. Знаеше руский, сръбский, румжнский, турский, гръцкий, старогрѣцкий, французский, арабский и дори отъ чясти санскрит-



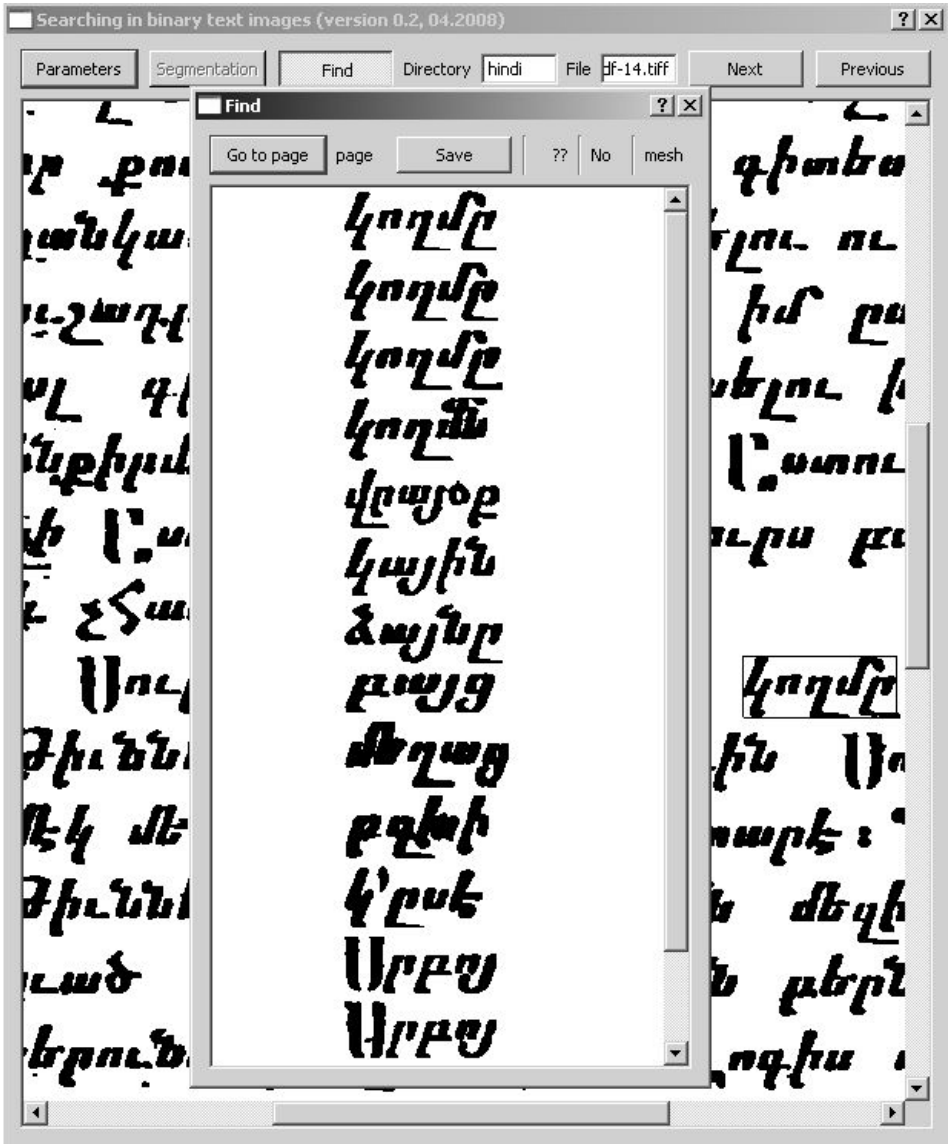
Old Greek text (approx. in the third century BC), 50 pages, jpg (1077×1416), 8 BPP

τοιαῦτα κατὰ σχῆμα πάντη εἰσὶν ἄτρεπτα, οἷα μ
Σῶμα δὲ σύνθετον ἐν τῇ φύσει ὐδὲν τοιοῦτον· ὥπ
θλίψει, τομῇ ὁπωσοῦν ὑπείκει, ᾗ λίθακες οἱ σερῤῥ
τητι διαφέρων ἀδάμας, μηδενὸς ὅλως ἐξαιρεμένου
μοιρον, ὡς δέδεικται. Ὧν δὲ σωμάτων τὰ μέρη
κει, ᾗ ῥᾷστα διαχωρίζεσθαι πέφυκεν ἁπαλὰ ταῦτ
μέλι, ἄργιλλος, κτ:· ὅσῳ δὲ ἧττονι ᾗ ἀτονωτέρᾳ
πλοκῆς ἀποσείχει, τοσούτῳ ᾗ ἁπαλώτερα, ἐς ἄκρ
ὐδενὶ γάρ ἐστιν ἐντυχεῖν, οὗ τὰ μέρη μὴ ὁπωσοῦν
γνύμενα.

Γ'. Τὸ σκληρὸν σῶμα ὑπὸ κώφητός τε ᾗ ἀοθεν
πεφυκὸς, εὔθραυστον ἀκούει· τοιαῦτα χάλυψ
τὰ κεράμεια σκεύη· Τούτων τὰ σερεὰ μέρη ὐχ ὅτω
ἐλλήλοις, διὸ ᾗ ῥᾷστα τῆς ἀμοιβαίας παφῆς ἀφίστατ
Δ'. Τὸ ἐκ πολλῶν οἱονεὶ λεπίδων πάνυ λεπτῶν ἀ
σάμενον, εὔσχιστον σῶμα προσείρηται· τούτου τὰ
μόρια σερῥότερον προσκεκόλληται ἀλλήλοις, ἢ λεπὶς
ἄρα τὰ τοιαῦτα σώματα εἰς λεπίδας ἀναλυόμενα·
καλούμενοι λίθοι οἱ ἐξ Ἰβηρίας, ᾗ Καππαδοκίας,



Armenian book (1858), 178 pages, tiff (2800×5000), 1 BPP



Text in Spainish (1901), 30 + 57 pages, gif (1400×2500), 4 BPP

Alonso, Rogelio M., Cartilla histyrico-descriptiva del tŭrmino municipal de Macuriges. Habana: Impr. La Propagandista, 1901, HOLLIS Catalog, Harvard University

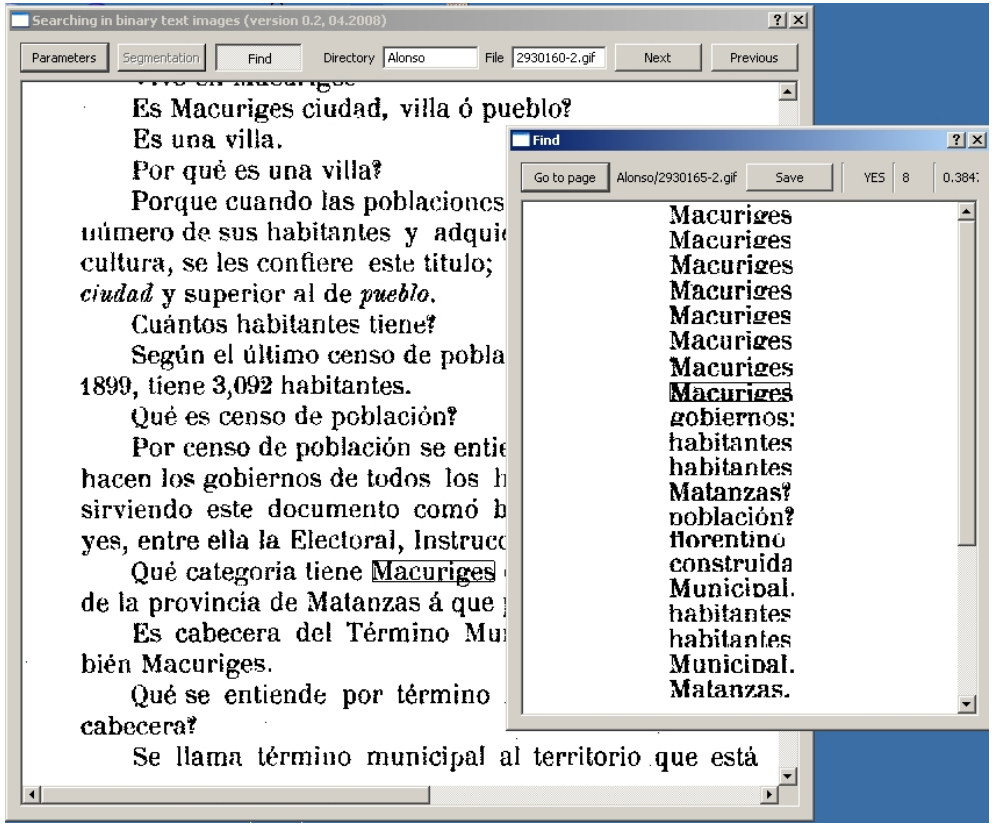
No señor; todas las fincas azucareras tienen sus chuchos que conectan con las líneas del ferrocarril y hay además caminos reales, trasversales y vecinales, estos en estado natural. (1)

Qué entiende V. por caminos reales, trasversales y vecinales?

Caminos reales, son los caminos abiertos por el gobierno Español desde los tiempos primeros de la colonización de la Isla de Cuba y tienen de ancho 24 varas; caminos trasversales son los que solo tienen de ancho 12 varas y vecinales los pasos permitidos por los propietarios de fincas, para acortar distancias de un lugar á otro y salvar lo mal que pudieran estar los caminos por el fango, las piedras ó la yerba.

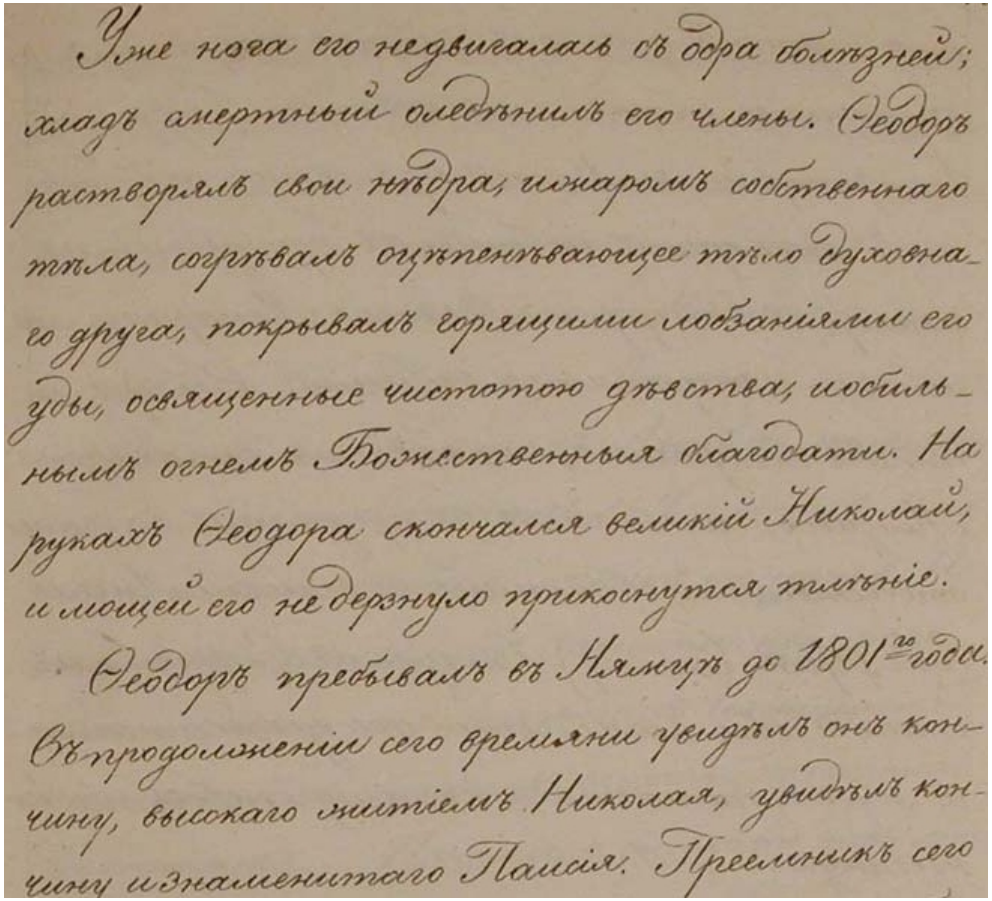
Cuántos ingenios para la fabricación de azúcar tiene en la actualidad el Término todo?

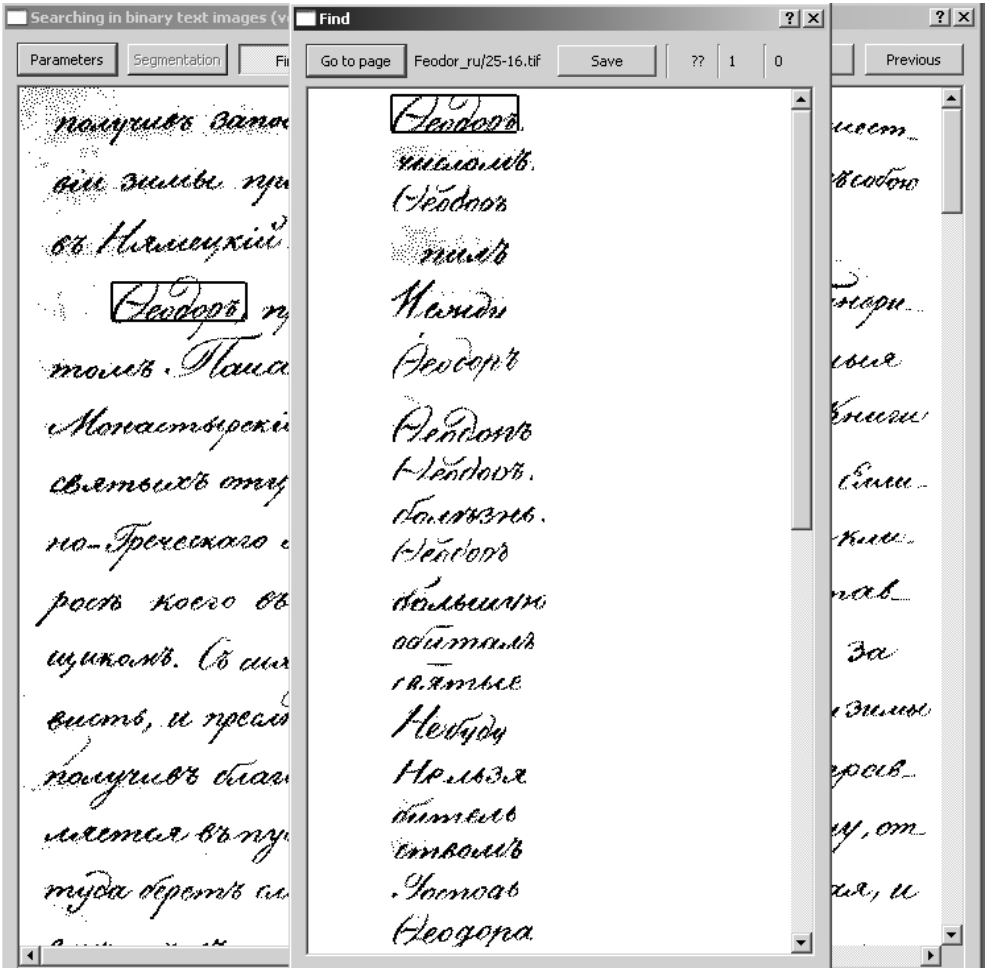
Los siguientes: «Santa Filomena» en el barrio de Navajas propiedad del Sr. Leandro Soler, «Elizalde» del Sr. Alberto Broch en el Ciego y «Santa Catalina» del señor Enrique Heedigg en el mismo barrio; «Carmen» del Sr. Alexander en Navajas, «Socorro» del Sr. Pedro Arenal en Tramojos y «Dolores» del Sr. Francisco Rosell en Platanal, todos centrales y con magníficos aparatos.



Handwritten document in Russian (1840), 44 pages, jpg (700×900), 24 BPP

Дом живоначальной Троицы, Свято-Троицкая Сергиева Лавра, Собрание славянских рукописей, 43: Житие схимонаха Феодора

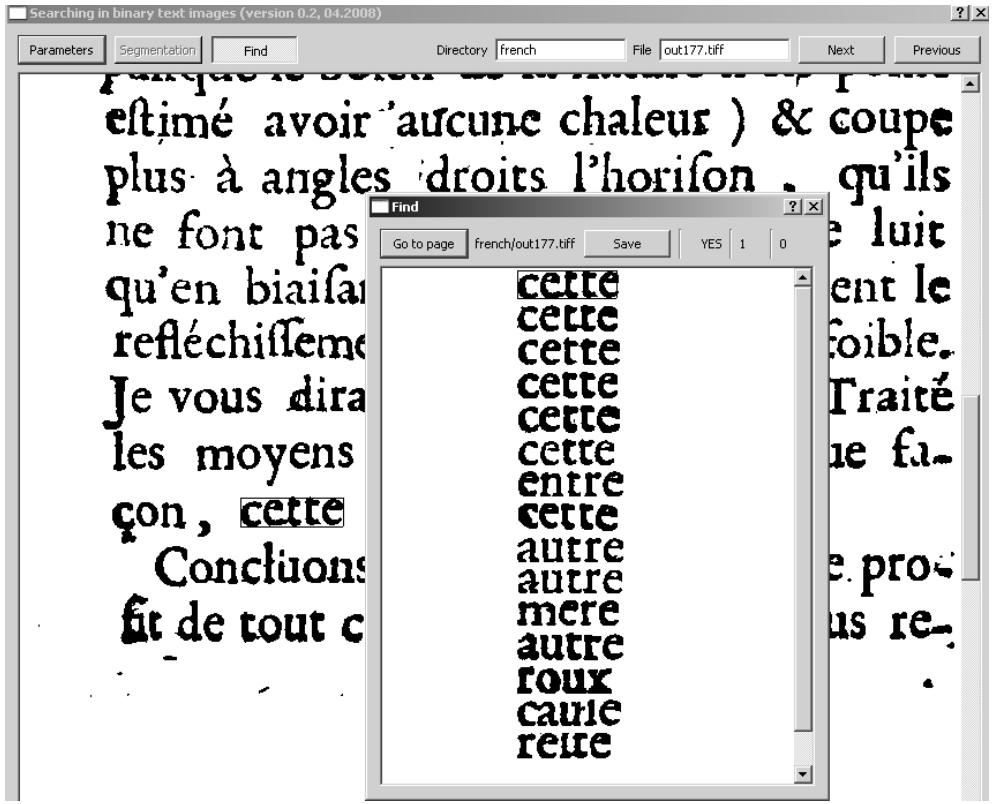




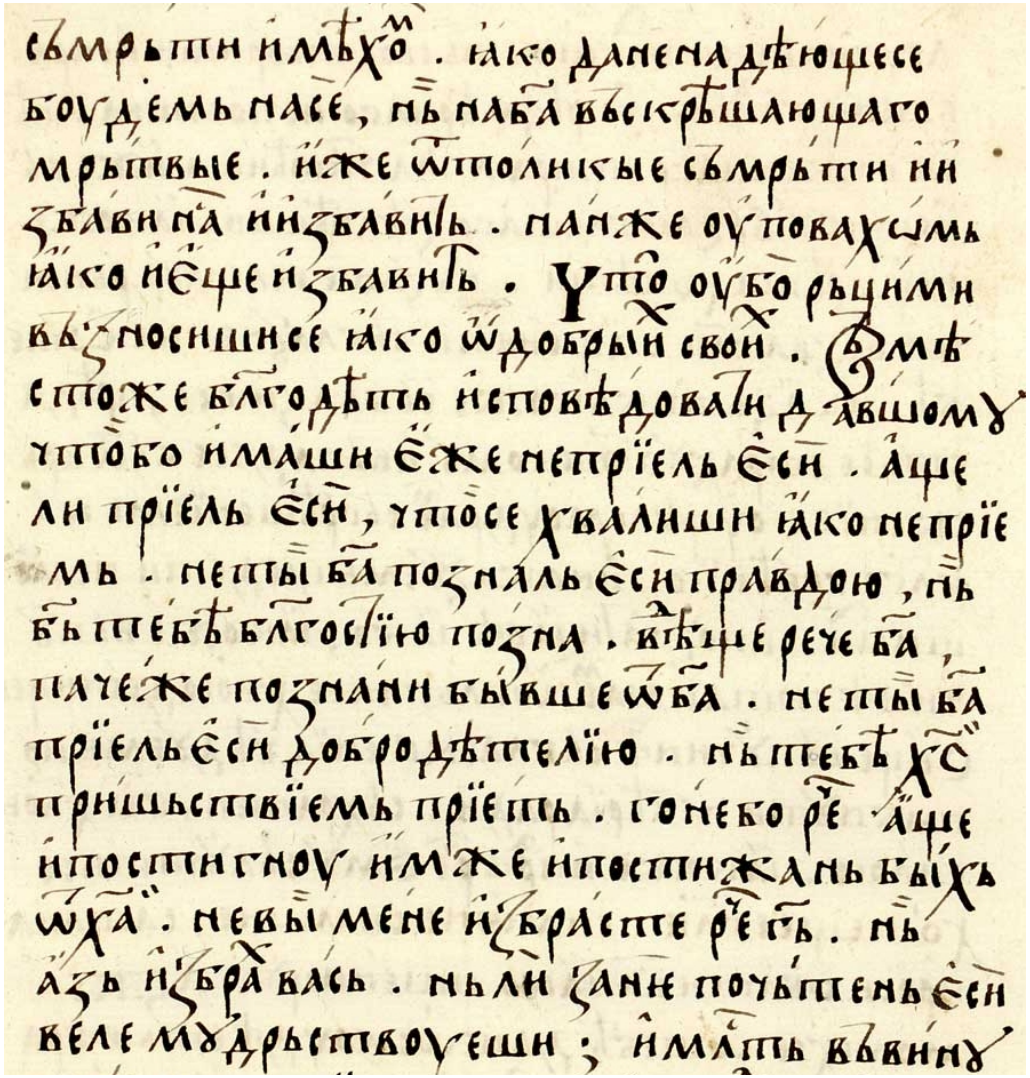
Text in French (1692), 388 pages, jpg (2048×3550), 8 BPP

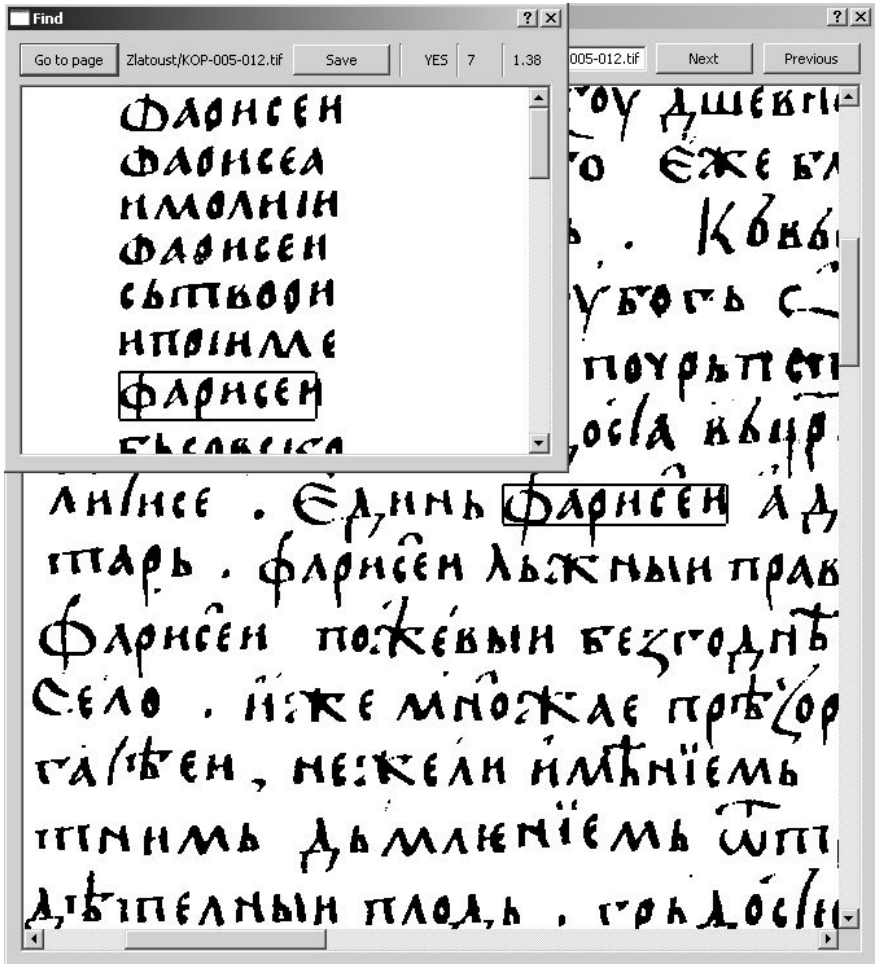
Nicolas de Bonnefons, Ch. de Sergy, (1692), University of Gent, Digitized by Google (2007)

Quant à la Terre, ſi vous la rencontrez bonne, ce vous ſera un grand avantage,& une grande épargne ; mais rarement en pourrez-vous trouvez,où il n'y ait beaucoup à travailler , dautant que telle paroîtra paſſablement bonne au deſſus, qui étant ouverte de la profondeur d'un fer de Béche ſeulement, ſe trouvera Argileuſe deſſous ; ce fonds eſt pire aux Arbres que le Tuf,ou la Roche,à cauſe qu'il s'y rencontre de petites veines où les Racines peuvent s'étendre & profonder,afin de tirer la fraîcheur de plus bas,& prendre quelqué nourriture ; mais l'Argileuſe ou Terrre franche ou rouge, fait comme un plancher qui par ſa dureté & denſité , ne peut être percé par aucunes Racines, & qui dans les grandes ardeurs de l'Eté,em-



Slavonic manuscript, (1574), 747 pages, jpg (1249×1890), 24 BPP.
Дигитална Народна библиотека Србије, Ћирилски рукописи, Збирка словенских рукописа Јернеја Копитара, Зборник “Златоуст”





Conclusion

We show the possibilities of preliminary version of the program.

Experiments with 7 different languages from various epochs give us the certainty in practical benefit of our approach.

It is quite universal and does not required any specific features of the concrete language.

Word searching can be applied to any collection of scanned documents, immediately after the graphic files have been created by the scanner device.

In spite of relatively low efficiency of the Hausdorff type methods (the searching process takes a lot of time) high level personal computers will be able to solve the problem in reasonable time.

We think also that the accuracy of retrieval is quite sufficient for practice.

There are some directions to improve the software.

- Increasing the efficiency and speeding up the search.
- Searching with a part of word as a pattern.
- Character segmentation of a page (or pages) and composing pattern word from well separated letters.
- Feedback – making second search for the same word with a different pattern word.
  - The user can choose this word among correct words found in the first search.
  - Produce a new pattern as an average of all or part of the correct words.
- Automatic or semi automatic choice of parameters based on image information.

Thank you for the attention.