Accessing the content of Greek historical documents

Anastasios Kesidis
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos", GR-15310,
Agia Paraskevi, Athens, Greece
+30 210 6503140
akesidis@iit.demokritos.gr

Eleni Galiotou
Department of Informatics
Technological Educational Institution of Athens
Ag. Spyridona, GR-12210 Egaleo, Athens, Greece
+30 210 5385805
egali@teiath.gr

Basilis Gatos
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos", GR-15310
Agia Paraskevi, Athens, Greece
+30 210 6503183
bgat@iit.demokritos.gr

Aristomenis Lampropoulos
Department of Informatics
University of Piraeus
Karaoli & Dimitriou 80, GR-18534 Piraeus, Greece
+30 210 4142263
arislamp@unipi.gr

Ioannis Pratikakis
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications,
National Center for Scientific Research "Demokritos", GR-15310
Agia Paraskevi, Athens, Greece
+30 210 6503183
ipratika@iit.demokritos.gr

Ioanna Manolessou
Angela Ralli
Department of Philology
University of Patras
University campus, GR-26504 Rio, Patras, Greece
+30 2610 969943
manolessou@upatras.gr
ralli@upatras.gr

ABSTRACT

In this paper, we propose an alternative method for accessing the content of Greek historical documents printed during the 17th and 18th centuries by searching words directly in digitized documents based on word spotting, without the use of an optical character recognition engine. We describe a methodology according to which synthetic word images are created from keywords. These images are compared to all the words in the digitized documents while user feedback is used in order to refine the search procedure. In order to improve the efficiency of accessing and searching, we have used natural language processing techniques that comprise (i) a morphological generator for early Modern Greek which provides the users with the ability to search documents using only a word stem and locate all the corresponding inflected word forms and (ii) a synonym dictionary which facilitates access to the semantic context of documents and enriches the results of the search process.

AND '09, July23-24, 2009, Barcelona, Spain.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing - Linguistic processing; I.5.4 [**Pattern Recognition**]: Applications - Text Processing; I.7.5 [**Document and Text Processing**]: Document Capture - Document Analysis; J.5 [**Computer Applications**]: Art and Humanities - Literature

General Terms

Algorithms; Design; Experimentation

Keywords

Historical document indexing, word spotting, natural language processing, computational morphology

1. INTRODUCTION

A robust indexing of historical printed documents is essential for quick and efficient content exploitation of the valuable historical collections. In this paper, we deal with Greek historical documents printed during the 17th and 18th centuries. Traditional approaches in document indexing, usually involve an Optical Character Recognition (OCR) step [3]. Usually, printed OCR systems involve a character segmentation step followed by a recognition step using pattern classification algorithms. In the case of printed historical documents OCR,

several factors affect the final performance like low paper quality, paper positioning variations (skew, translations, etc), low print contrast, typesetting imperfections, etc. In literature, two general approaches can be identified: the segmentation approach and the holistic or segmentation-free approach. The segmentation approach requires that each word has to be segmented into characters while the holistic approach entails the recognition of the whole word. In the segmentation approach, the crucial step is to split a scanned bitmap image of a document into individual characters [6]. A holistic approach is followed in [8], [11], [13], [14], [16], where line and word segmentation is used for creating an index based on word matching.

Accessing the content of the documents necessarily implies the use of natural language processing (NLP) techniques. In particular, inflected word forms which are used in the word spotting procedure are usually produced by a morphological generation tool. For the implementation of the morphological generator finite state techniques proved to be quite successful in many NLP applications covering a wide spectrum of languages [2], [10], [18], [21]. Furthermore, access to the semantic content of the documents is facilitated by the implementation of a synonym dictionary [25].

In this paper, we propose an alternative method for accessing the content of Greek historical documents printed during the 17th and 18th centuries by searching words directly in digitized documents based on word spotting aided by NLP techniques, without the use of an optical character recognition engine.

The paper is organized as follows: Section 2 describes the general framework of the proposed system. Section 3 concerns the word spotting process and describes its several phases. Section 4 describes the linguistic and computational approaches for the morphological generation system while section 5 details the synonym dictionary. Experimental results that demonstrate the proposed method are given in section 6. Finally, conclusions are drawn in section 7.

2. THE PROPOSED FRAMEWORK

Our work concerns a collection of digitized Greek documents printed during the 17th and 18th centuries. These texts contain a specific morphology, which reflects the evolution of the Greek language through the particular time period containing word-forms from Ancient, Medieval and Modern Greek. The proposed framework can be part of an information system for the processing, management and accessing to the content of the aforementioned collection of historical books and manuscripts. Figure 1 depicts the overall architecture of the proposed system. The word spotting component of the system introduces a novel way to initialize the word retrieval mechanism through the creation of synthetic word data along with a robust hybrid feature extraction that supports meaningful representations of word images. It consists of the following two phases:

i. A preprocessing step that aims to improve the quality of the image followed by robust word segmentation. Several image operations are applied including binarization, enhancement, orientation and skew correction, noisy border and frame removal. Then, a segmentation process follows, that locates the words in the document.

ii. A two-step word retrieval phase that aims to rank the segmented words according to their similarity to the query word. In this phase, user feedback is applied that improves drastically the matching results.

The aim of the natural language system component is the use of advanced natural language processing techniques for the intelligent and effective searching in the collection such as the word-spotting procedure. Query extension necessarily implies:

i. a morphological generation system which provides users the ability to search documents using only a word stem and locate all the corresponding inflected words.

ii. a synonym dictionary which facilitates the access to the semantic context of documents and enriches the results of the search process.

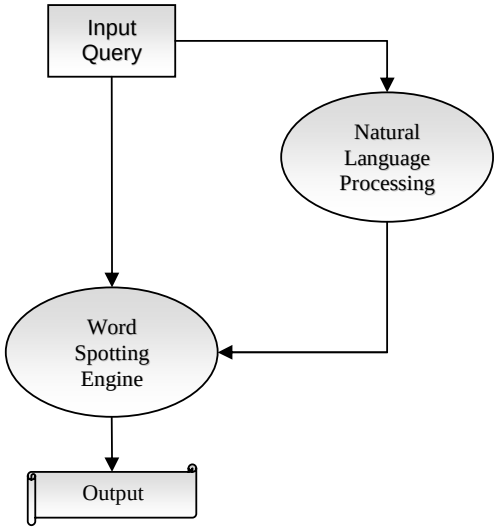


Figure 1. System architecture

3. WORD SPOTTING

In the case of historical documents, Manmatha and Croft [15] presented a holistic method for word spotting wherein matching was based on the comparison of entire words rather than individual characters. Their method involved an off-line grouping of words appearing in a historical document and the manual characterization of each group by the ASCII equivalence of the corresponding words. This process can become very tedious when processing large collections of documents. In order to eliminate this process, we use a method for keyword search in historical printed documents which is based on [14] and consists of the following steps: (i) image preprocessing; (ii) creation of synthetic image words from keywords typed by the user; (iii) word segmentation using dynamic parameters; (iv) efficient feature extraction for each image word and (v) a retrieval procedure that is supported by user's feedback. The word spotting procedure is initialized by applying the synthetic keyword image as the query image for the retrieval of all

relevant words. The synthetic word is matched against all detected words and the accuracy of the results is further optimized by the user's feedback. Combination of synthetic data creation and user's feedback leads to improved results in terms of precision and recall. The proposed workflow for keyword searching in historical printed documents is presented in the sequel.

3.1 Image Preprocessing

To aid towards a correct segmentation of historical documents into words we follow three distinct steps, namely image binarization and enhancement, orientation and skew correction and noisy border and frame removal. A detailed description of each step is given in the sequel.

3.1.1 Image Binarization and Enhancement

Binarization refers to the conversion of the gray-scale image to a binary image and is the starting step of most document image analysis systems. Moreover, since historical document collections are usually characterized by very low quality, an image enhancement stage is also essential. The proposed scheme for image binarization and enhancement is based on the work of [7] and consists of five distinct steps: a pre-processing procedure using a low-pass Wiener filter, a rough estimation of foreground regions, a background surface calculation by interpolating neighboring background intensities, a thresholding by combining the calculated background surface with the original image and finally a post-processing step that improves the quality of text regions and preserves stroke connectivity. The last step also eliminates noise and improves the quality of text regions by removing isolated pixels and filling possible breaks, gaps or holes.

3.1.2 Orientation and skew correction

It is necessary to identify and correct the text orientation (portrait or landscape) and skew before proceeding to the word segmentation phase. Text orientation is determined by applying a horizontal/vertical smoothing, followed by a calculation procedure of vertical/horizontal black and white transitions [28]. For skew detection, a fast Hough transform approach is used that is based on the description of binary images using rectangular blocks [17].

3.1.3 Noisy Border and Frame removal

Often, images resulting from document scanning are framed by a solid or stripped black border. The methodology used for border removal is based on projection profiles combined with a connected component labeling process while signal cross-correlation is also used in order to verify the detected noisy text areas [23]. Additionally, image projections are used at the deskewed image in order to remove noisy text from neighboring pages. Finally, in order to ease the segmentation process any potential frames around the text areas are also removed. Lines are detected by processing horizontal and vertical black runs as well as by morphological operations with suitable structuring elements that connect possible line breaks and enhance the line segments [5].

3.2 Segmentation

In the proposed methodology segmentation is accomplished with the use of the Run Length Smoothing Algorithm (RLSA) [24], [27] by using dynamic parameters which depend on the average character height as described in [13]. RLSA examines the white runs existing in the horizontal and vertical directions. For each direction, white runs with length less than a threshold are eliminated. In the proposed method, the horizontal length threshold is experimentally defined as 50% of the average character height while the vertical length threshold is experimentally defined as 10% of the average character height. The application of RLSA results in a binary image where characters of the same word become connected to a single connected component. In the sequel, a connected component analysis is applied using constraints which express the minimum expected word length. This allows the reject of stop-words and therefore eliminates undesired word segmentation.

3.3 Synthetic Data Creation

The creation of synthetic image words from keywords typed by the user refers to the artificial creation of the keyword images from their ASCII equivalences. Thus, an image template for all required characters has to be previously defined and stored to the system. As shown in Figure 2, during the manual character template marking, an adjustment of the baseline for each character image template is supported in order to correctly align the character templates when constructing the synthetic word image. We follow this procedure in order to avoid inaccurate alignment results since there are characters whose lower end is below the text line, e.g. the letter «ρ» in Figure 2. The character template selection is a process that is performed "once-for-all" and can be used for entire books or collections that share a common typewriting style.

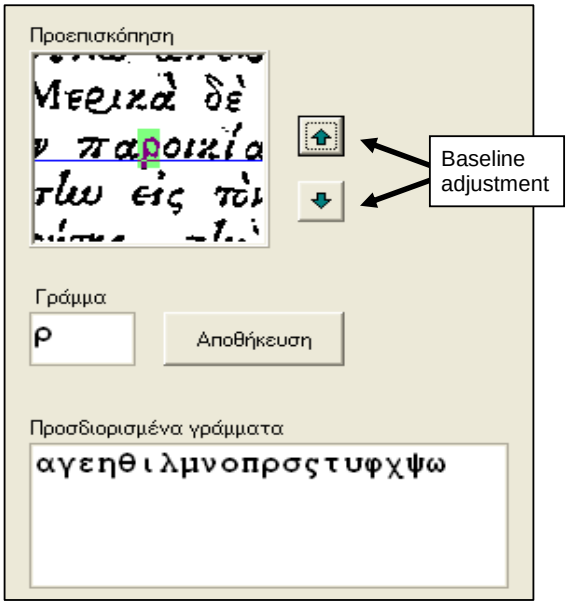


Figure 2. Marking the template for Greek character «ρ»

3.4 Feature extraction

The segmented words of the historical document as well as the synthetic query word are described by features which are used during the word retrieval phase in order to measure similarity

between word images. The feature extraction phase consists of two distinct steps; (i) normalization and (ii) hybrid feature extraction.

3.4.1 Normalization

The normalization process is dedicated to preserve scale invariance for word images. In particular, for the normalization of the segmented words we use a bounding box with user-defined dimensions. The segmented words are resized to fit in the bounding box while preserving their aspect ratio. The size of the bounding box concerning both width and height is the same for all words. Thereafter, exact positioning of the word in the bounding box is achieved by placing the geometric center of the word in the center of the bounding box.

3.4.2 Hybrid feature extraction

Several features and methods have been proposed for word image matching based on strokes, contour analysis, etc [4], [19]. In the proposed approach, two different types of features are used. The first one, which is based on [3], divides the word image into a set of zones and calculates the density of the character pixels in each zone. The second type of features is based on word (upper/lower) profile projections. The word image is divided into two sections with respect to the horizontal line which passes through the center of mass (x_t, y_t) of the word image. Upper/lower word profiles are computed by recording, for each image column, the distance from the upper/lower boundary of the word image to the closest character pixel [19]. The word is divided into y_w vertical zones. For each upper/lower word profile we calculate the area in the upper and lower sections that correspond to the desired features. We build a set of 90 features based on zones and a set of 60 features based on profile projections, leading to an overall of 150 features for the hybrid scheme.

3.5 Word matching

The process of word matching involves the comparison/matching between the synthetic query word image and all the indexed segmented words. Ranking of the comparison results is based on calculating the L_1 distance metric using the above motioned features. The selection of the L_1 distance metric is based on experimentations that have shown its effectiveness in accordance with low computational cost. An initial list of results is produced that sorts the segmented word images of the document according to their similarity to the synthetic word image created by the user. These results are refined in the user's feedback phase.

3.6 User feedback

From the initial ranking we obtained the words of the document that are similar to the synthetic keyword. These results might not present high accuracy because a synthetic keyword cannot a priori perform a perfect match with a real word image. User feedback is an efficient mechanism for drastically improving the matching process. We propose a user intervention where the user selects as input query one or more correct results from the list produced after the initial word matching process. Then, a new matching process is initiated. The segmented words are ranked according to their similarity to the selected word(s) which, in this case, are not synthetic but real words of the document's corpus. The critical impact of the user feedback in the word spotting process lies upon this transition from synthetic to real data. Furthermore, in the proposed system the user interaction is supported by a simplified and user friendly graphical interface that makes the word selection procedure an easy task. Figure 3 illustrates the flow diagram of the user feedback process.

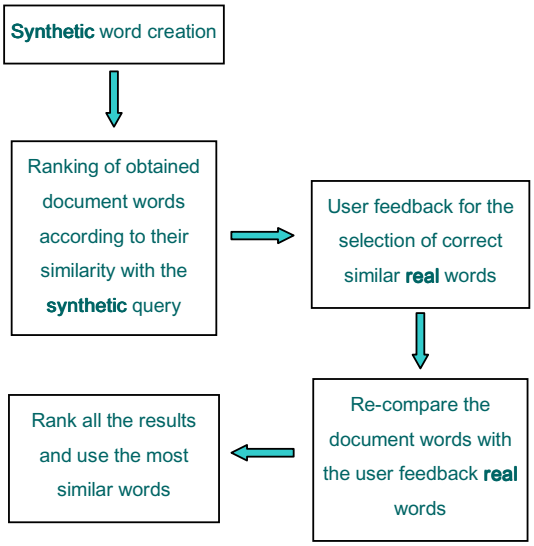


Figure 3. The consecutive stages of the user feedback process

4. A MORPHOLOGICAL GENERATOR FOR EARLY MODERN GREEK

4.1 The Linguistic Approach

Our experimentation is performed on a corpus which consists of digitized historical and religious documents printed during the 17th and 18th centuries. The language of the particular time period reflects an early stage of Modern Greek and represents the evolution of the Greek language since it incorporates elements from Ancient, Medieval and Modern Greek. We have chosen a set of keywords which entail historical, theological, philological and linguistic interest. These words are mainly nouns of relevant high frequency of appearance which characterize the documents. We also took into account some grammatical elements of particular linguistic interest. These keywords were further analyzed into their morpheme constituents and morphologically characterized. In particular, the task of the morphological processing of the selected word-forms was divided into the following subtasks:

i. Grammatical categorization of word-forms and classification into grammatical categories (nouns, adjectives, adverbs etc)

ii. Recognition of the internal morpheme components of the words

iii. Grammatical characterization of morphemes and assignment of morpho-syntactic information per category (e.g. number, case, gender etc)

iv. Analysis of the words into stems and affixes

v. Writing of rules to be used for the computational morphological processing of word-forms

The results of the abovementioned subtasks constitute the linguistic basis of our morphological processing tool which will be described in the following subsection.

4.2 The Computational Approach

4.2.1 The Finite State Approach to Computational Morphology

In recent years, finite state techniques have been widely used in many NLP applications such as tagging, parsing, information extraction [10]. Finite state approaches to morphology proved to be particularly successful covering a large spectrum of languages [12], [9], [1]. Morphological phenomena in Modern Greek have been extensively dealt with, within the finite state framework covering inflection, derivation and compounding [18], [22]. Therefore, finite state technology was a promising choice for the implementation of our computational morphology that reflects a transitional period in the evolution of the Greek language and has not yet been systematically analyzed.

The finite-state approach to morphology is based on the representation of the relation between the surface forms of a language and their corresponding lexical forms. This relation can be described as a regular relation using the metalanguage of regular expressions and, with a suitable compiler the regular expression source code can be compiled into a finite-state transducer that implements the relation computationally. As a result, a transducer represents a mapping between a surface form and the lexical form through a sequence of states and arcs from an initial state to a final one.

4.2.2 SFST-based Morphological Processing

We have developed a software tool for the creation, manipulation and application of computational morphologies which embeds the SFST tools [21], [20] - a collection of software tools for the generation, manipulation and processing of finite-state transducers which are specified by means of the SFST programming language. A SFST program is essentially a regular expression. The SFST system was developed by the Institute for Natural Language Processing, University of Stuttgart. It comprises a compiler, which translates finite state transducer programs (regular expressions) into minimized transducers and a wide range of transducer operations similar to commercial platforms such as XFST [2]. We opted for the SFST system because it is an open source software and it supports UTF-8 character coding which is important for the implementation of Greek computational morphologies. The source code of the SFST-PL (SFST Programming Language) can be downloaded from the url: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/SFST/. It has been compiled using the MinGW and MSYS open software that can be downloaded from the url: http:/www.mingw.org/ and served as the basis of the SFST compiler.

4.3 The Morphological Generator

4.3.1 The system architecture

The architecture of the morphological generation tool is illustrated in Figure 4.

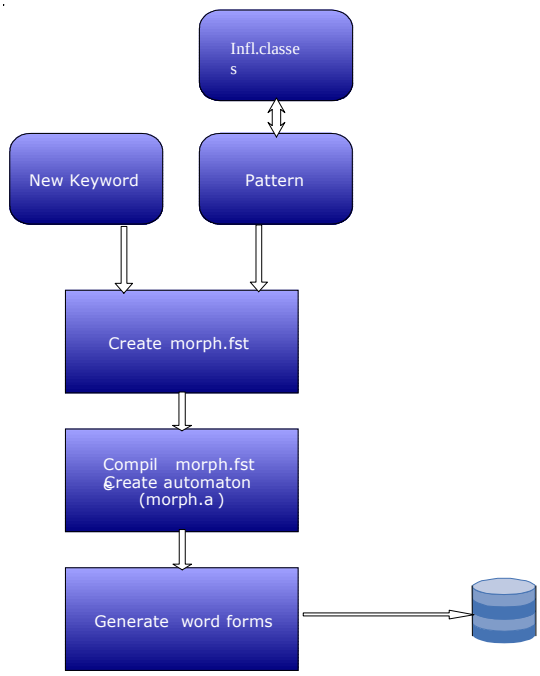


Figure 4. Architecture of the Morphological Generator

The system provides the user with the possibility to insert a new keyword and select a pattern as a representative of the appropriate inflectional class. Then, the file *Morph.fst* is created and dynamically compiled into the *morph.a* automaton that is responsible for the generation of word-forms. Figure 5 illustrates the inflectional classes' definition in our system.

4.3.2 The morphological generation process

We have applied a lexeme-based approach to morphological processing where a word-form is considered as the result of applying rules that alter a word-form or stem in order to produce a new one. An inflectional rule takes a stem, changes it as is required by the rule and outputs a word-form. During the generation process users are able to:

i. insert a new keyword

ii. assign the appropriate inflectional class through the selection of the appropriate representative

iii. perform a generation of all inflected word-forms according to the function of the selected representative of the inflectional class as illustrated in Figure 6.

At this stage of the processing the user needs not to be familiar with the finite state formalism. Moreover, no special linguistic knowledge is required by the user in order to choose the

appropriate representative according to which the generation process will take place. Knowledge of the language at the native speaker level will suffice.

As a result to this procedure the system enriches dynamically the list of keywords and inflected word-forms are used in the word spotting procedure.

```
%%% definition of the inflectional classes %%%
$1a_1$= {<nom-sing>}:{ας} |\
   %{<nom-sing>}:{αν} |\
   {<gen-sing>}:{ου} |\
   %{<voc-sing>}:{α} |\
   {<acc-sing>}:{α} |\
   {<acc-sing>}:{αν} |\
   {<nom-pl>}:{αι} |\
   {<gen-pl>}:{ων} |\
   %{<voc-pl>}:{αι} |\
   {<acc-pl>}:{ας}

$1a_2$= {<nom-sing>}:{ης} |\
   %{<nom-sing>}:{αν} |\
   {<gen-sing>}:{ου} |\
   %{<voc-sing>}:{α} |\
   {<acc-sing>}:{η} |\
   {<acc-sing>}:{ην} |\
   {<nom-pl>}:{αι} |\
   {<gen-pl>}:{ων} |\
   %{<voc-pl>}:{αι} |\
   {<acc-pl>}:{ας}

$1b_1$= {<nom-sing>}:{ας} |\
    {<acc-sing>}:{α} |\
   {<acc-sing>}:{αν}

$1b_2$={<nom-pl>}:{ες} |\
     {<gen-pl>}:{ων}

$1g$= {<nom-sing>}:{αν} |\
   %{<nom-sing>}:{α} |\
```

Figure 5. Inflectional classes definition

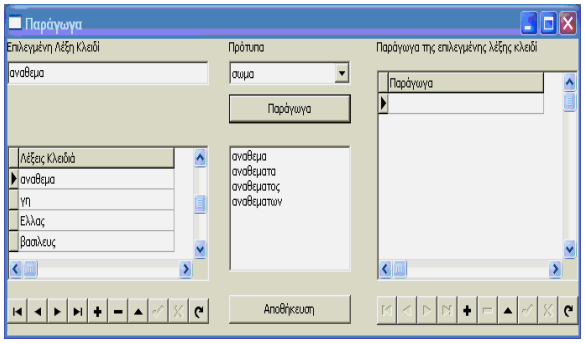


Figure 6. Morphological Generation Process example. The keyword is «αναθεμα», the inflectional class representative is «σωμα» and the inflected word-forms are «αναθεμα, αναθεματα, αναθεματος, αναθεματων»

5. THE SYNONYM DICTIONARY

In order to facilitate the access to the semantic context of documents and to enrich the results during the search process, the use of semantic networks such as WordNet is suggested [26]. Yet, in order to improve the efficiency of our system we have embedded a synonymy based on a relevance feedback technique since its functionality is equivalent to that of a semantic network and at the same time it avoids an unnecessary amount of ambiguity that could affect the accuracy and efficiency of the system [25]. Our approach is based on the methodology proposed in [25] for adapting a synonym database to specific domain. Our synonym dictionary provides the user with two options:

i. Add a word in the dictionary accompanied by up to five synonyms. Each synonym is given a weighted value of the relevance of the synonym to the word added. These values range from 1 to 10, the highest relevance to the word in question corresponding to weighted value of 1. Additionally, the options of editing or deleting words from the dictionary are available.

ii. Look up a word and its synonyms. With the use of a graphical user interface the user can look up a word in the dictionary as well as its meaning and its weighted synonyms. The ensemble of words found (original word and the corresponding synonyms) can be used to update the query in the word spotting process.

Both the results of the morphological generator process as well as the synonym dictionary can be used for updating the query for the word spotting in order to enrich the search results.

6. EXPERIMENTAL RESULTS

The experimental results regarding the word spotting procedure were based in a corpus of 12 typewritten pages in which a keyword has been searched. Initially, the keyword has been synthetically created from the character templates that correspond to their ASCII equivalences. The synthetic keyword has been applied as the query image for the retrieval of all relevant words. The first ranked results have been used in order to select the correct words of the corpus according to the user feedback performing a transition from synthetic to real data. Then, a new matching process is initiated. The segmented words are ranked according to their similarity to the selected real words of the document's corpus. Figure 7 illustrates the matching results before and after the user feedback process.

The effect of user's feedback in increasing the word spotting performance has been evaluated by varying the number of user selected words in the second phase of the process. Specifically, the final ranking results have been examined when the user selects two, three and five real words, respectively (see Table 1). It can be clearly seen that increasing the number of selected real words improves the efficiency of the word spotting procedure by resulting in more correct words at the first twenty entries of the ranking list.

We have also implemented and tested the morpho-phonological rules that cover the nominal inflection phenomenon. To this end, we have defined 14 inflectional classes, 9 of which are considered as the major ones. A fragment of the inflectional classes file appears in Figure 5. We have processed some 126 words which characterize our corpus thus producing more than 500 inflected word-forms. In this way, the word spotting procedure was able to locate all inflected word-forms of a candidate base form. Moreover the user was able to extend his query with the use of selected synonyms of the keywords from the synonym database. Figure 8 shows the extraction of inflected word forms and synonyms for a given word and their use in the search process.

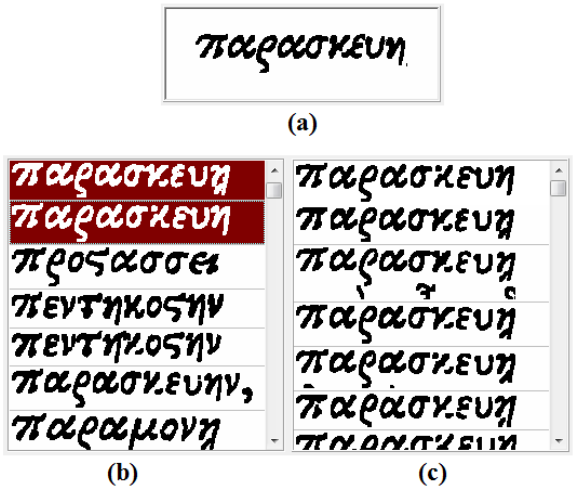


Figure 7. (a) Synthetic query word (b) Initial ranking of segmented words. The highlighted words denote correct words selected by the user (c) Ranking after user's feedback.

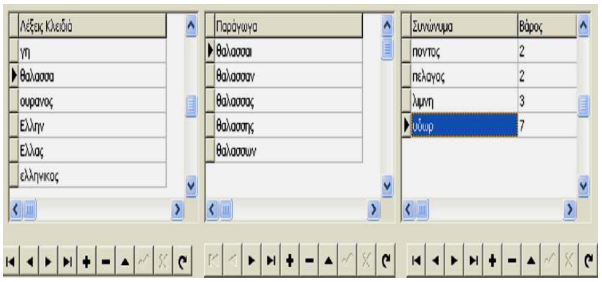
Table 1. Ranking results for different number of user selected words

Selected correct words	Query word instances in the ranking results	Percentage of correct words in the first 20 results
2	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 24, 36, 42	81.2%
3	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 25, 37, 43	81.2%
5	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 22, 28	87.5%

Emphasis was placed on the extensibility of the system as for the enrichment with new keywords and the generation of their corresponding word-forms. The system can be enriched with new words simply by defining the correspondence to the words that are representative of the inflectional classes. In addition, the generator was designed so as to be able to incorporate more morphological elements and deal with the phenomena of derivation and compounding. The choice of the meta-language for the construction of the automata facilitates the experimentation with other linguistic phenomena without any substantial change to the system and provides a systematic and standard way for the implementation of regular relations.

7. CONCLUSIONS

In this paper, we have described a methodology for accessing the content of digitized Greek historical documents without the use of an optical character recognition system. Access to the documents which were printed during the 17th and 18th centuries is based on word spotting using word images created from keywords.



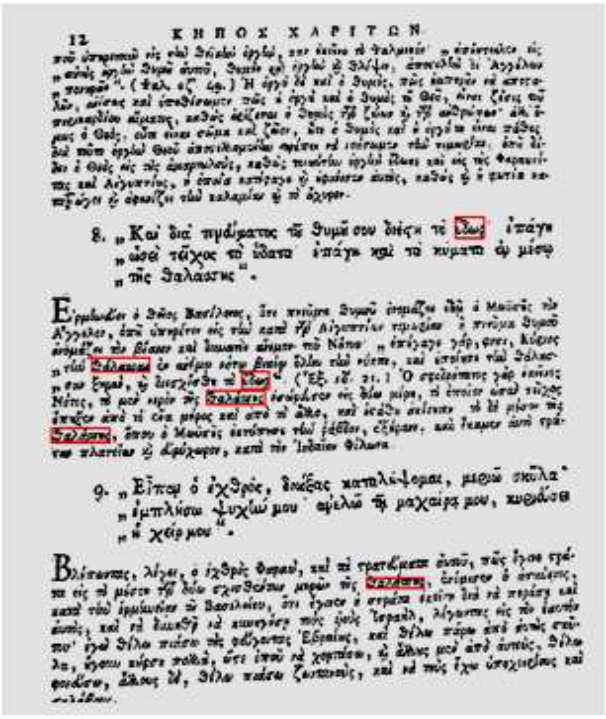


Figure 8. (a) Extraction of inflected word forms and synonyms («υδωρ») for given word («θαλασσα»); (b) use of extracted words in the word spotting process.

By using a method of creating synthetic word images from their ASCII equivalences we achieved word search while avoiding the process of character recognition that occurs in traditional indexing systems based on OCR. The user is able to further improve the results by selecting the most relevant words from the list of the results and feed them back to the system. User's feedback enables the transition from synthetic queries to real data queries resulting to more accurate searching results. Moreover, query extension and access to the semantic content of the documents are achieved with the use of a synonym dictionary and a morphological generation system. The morphological generator constitutes an attempt – the first to our knowledge - to build a morphological processor able to deal

with nominal inflection phenomena in early Modern Greek which represents a transitional period of the language.

8. REFERENCES

[1] Antworth, E., 1990. PC-KIMMO: A Two-level Processor for Morphological Analysis, Occasional Publications in Academic Computing no 16, Summer Institute of Linguistics, Dallas TX.

[2] Beesley, K., Karttunen, L., 2003. Finite State Morphology. CSLI Publications.

[3] Bokser, M., 1992. Omnidocument technologies, Proc. of the IEEE, 80(7), 1066-1078.

[4] Doerman, D., 1997. The detection of duplicates in document image databases, Proc. of the 4th Int. Conf. on Document Analysis and Recognition (ICDAR'97), 314-318.

[5] Gatos, B., Danatsas, D., Pratikakis I., Perantonis, S.J.: 2005. Automatic table detection in document images, Proceedings of the Third International Conference on Advances in Pattern Recognition (ICAPR'05), Lecture Notes in Computer Science (3686), 609–618.

[6] Gatos, B., Papamarkos, N., Chamzas, C., 1997. A binary tree based OCR technique for machine printed characters, Engineering Applications of Artificial Intelligence, 10(4), 403-412.

[7] Gatos, B., Pratikakis, I., Perantonis, S.J. 2006. Adaptive Degraded Document Image Binarization, Pattern Recognition, vol. 39, 317-327.

[8] Guillevic, D., Suen, C.Y., 1997. HMM word recognition engine, Fourth International Conference on Document Analysis and Recognition (ICDAR'97), 544–547.

[9] Karttunen, L., 1983. KIMMO: A General Morphological Processor, Texas Linguistic Forum, vol. 22, 163-186.

[10] Karttunen, L., Oflazer, K., 2000. Special Issue on Finite-State Methods in NLP: Computational Linguistics, vol. 26, no. 1.

[11] Keaton, P., Greenspan, H., Goodman, R., 1997. Keyword spotting for cursive document retrieval, Workshop on Document Image Analysis (DIA 1997), 74–82.

[12] Koskenniemi, K., 1983. Two-level Morphology: A General Computational Model for Wordform Recognition and Production, Publication No 11, Dept. of General Linguistics, University of Helsinki.

[13] Konidaris, T., Gatos, B., Ntzios, K.. Pratikakis I., Theodoridis, S., Perantonis, S. J., 2007. Keyword-Guided Word Spotting in Historical Printed Documents Using Synthetic Data and User feedback, International Journal on Document Analysis and Recognition (IJDAR), special issue on historical documents, Vol. 9, No. 2-4, 167-177.

[14] Lu, Y., Tan, C., Weihua, H., Fan, L., 2001. An approach to word image matching based on weighted Hausdorff distance, Sixth International Conference on Document Analysis and Recognition (ICDAR'01), 10–13

[15] Manmatha R., Croft, W.B., 1997. A Draft of Word Spotting: Indexing Handwritten Manuscripts, Intelligent Multimedia Information Retrieval, MIT Press, Cambridge, MA, 43-64.

[16] Marcolino, A., Ramos, V., Ármalo, M., Pinto, J.C., 2000. Lineand Word matching in old documents, Proceedings of the Fifth Ibero-American Symposium on Pattern Recognition (SIAPR'00), 123–125.

[17] Perantonis, S.J., Gatos, B., Papamarkos, N., 1999. Block decomposition and segmentation for fast Hough transform evaluation, Pattern Recognition, vol. 32(5), pp. 811- 824.

[18] Ralli, A., Galiotou, E., 2004. Greek Compounds: A Challenging Case for the Parsing Techniques of PC-KIMMO v.2, International Journal of Computational Intelligence, vol. 1, no. 2, 152-162.

[19] Rath T.M., Manmatha, R., 2003. Features for word spotting in historical documents, Proc. of the 7th Int. Conf. on Document Analysis and Recognition (ICDAR'03), 218-222.

[20] Schmid, H., 2005. A Programming Language for Finite State Transducers, Proc. FSMNLP 2005, Helsinki, Finland.

[21] Schmid, H., Fitschen, A., Heid, U., 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection, Proc. LREC 2004, Lisbon, Portugal, 1263-1266.

[22] Sgarbas, K., Kokkinakis, N. G., 1995. A PC-KIMMO-Based Morphological Description of Modern Greek, Literary and Linguistic Computing , 10(3), 189-201.

[23] Stamatopoulos, N., Gatos, B., Kesidis, A., 2007. Automatic Borders Detection of Camera Document Images, 2nd International Workshop on Camera-Based Document Analysis and Recognition (CBDAR'07), Curitiba, Brazil, 71-78.

[24] Theodoridis, S., Koutroumbas, K. 1997. Pattern recognition. Academic Press, New York.

[25] Turcato, D., Popowich, F., Toole, J., Fass, D., Nicholson, D., Tisher, D., 2000. Adapting a synonym database to specific domains, J. Klavans and J. Gonzalo J. (eds.) Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval, 1-11.

[26] Voorhees, E. M., 1998. Using WordNet for text retrieval, C. Fellbaum, (ed.) Wordnet: An Electronic Lexical Database, MIT Press Books, chap. 12, 285-303.

[27] Wahl, F.M., Wong, K.Y., Casey, R.G., 1982. Block segmentation and text extraction in mixed text/image documents, Comput. Graph. Image Process. 20, 375–390

[28] Yin, P.Y. 2001. Skew detection and block classification of printed documents, Image and Vision Computing 19, 567-579.