IJDAR (2011) 14:131–144
DOI 10.1007/s10032-010-0134-4

ORIGINAL PAPER

A word spotting framework for historical machine-printed documents

A. L. Kesidis · E. Galiotou · B. Gatos · I. Pratikakis

Received: 10 January 2010 / Revised: 28 July 2010 / Accepted: 12 October 2010 / Published online: 17 November 2010

Abstract In this paper, we propose a word spotting framework for accessing the content of historical machine-printed documents without the use of an optical character recognition engine. A preprocessing step is performed in order to improve the quality of the document images, while word segmentation is accomplished with the use of two complementary segmentation methodologies. In the proposed methodology, synthetic word images are created from keywords, and these images are compared to all the words in the digitized documents. A user feedback process is used in order to refine the search procedure. The methodology has been evaluated in early Modern Greek documents printed during the seventeenth and eighteenth century. In order to improve the efficiency of accessing and search, natural language processing techniques have been addressed that comprise a morphological generator that enables searching in documents using only a base word-form for locating all the corresponding inflected word-forms and a synonym dictionary that further facilitates access to the semantic context of documents.

Keywords Historical document indexing · Word spotting · Natural language processing · Computational morphology

A. L. Kesidis (✉) · B. Gatos · I. Pratikakis
Computational Intelligence Laboratory, Institute of Informatics and Telecommunications, National Center for Scientific Research "Demokritos", 15310 Agia Paraskevi, Athens, Greece
e-mail: akesidis@iit.demokritos.gr

B. Gatos
e-mail: bgat@iit.demokritos.gr

A. L. Kesidis
Department of Surveying Engineering, Technological Educational Institution of Athens, Ag. Spyridona, 12210 Egaleo, Athens, Greece
e-mail: akesidis@teiath.gr

E. Galiotou
Department of Informatics, Technological Educational Institution of Athens, Ag. Spyridona, 12210 Egaleo, Athens, Greece
e-mail: egali@teiath.gr

I. Pratikakis
Department of Electrical and Computer Engineering, Democritus University of Thrace, 67100 Xanthi, Greece
e-mail: ipratika@ee.duth.gr; ipratika@iit.demokritos.gr

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 religious documents printed during the seventeenth and eighteenth century. Traditional approaches in document indexing usually involve an optical character recognition (OCR) step [5]. 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, and typesetting imperfections. 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 [13]. A holistic approach is followed in [15,18,21,24,27], 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 that 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 [1,4,16,17,19,29,31,35]. Furthermore, access to the semantic content of the documents is facilitated by the implementation of a synonym dictionary [38].

In this paper, we propose an alternative method for accessing the content of Greek historical documents printed during the seventeenth and eighteenth century 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: Sect. 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 Sect. 5 details the synonym dictionary. Experimental results that demonstrate the proposed method are given in Sect. 6. Finally, conclusions are drawn in Sect. 7.

2 The proposed framework

Our work concerns a collection of digitized Greek documents printed during the seventeenth and eighteenth 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, which 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.

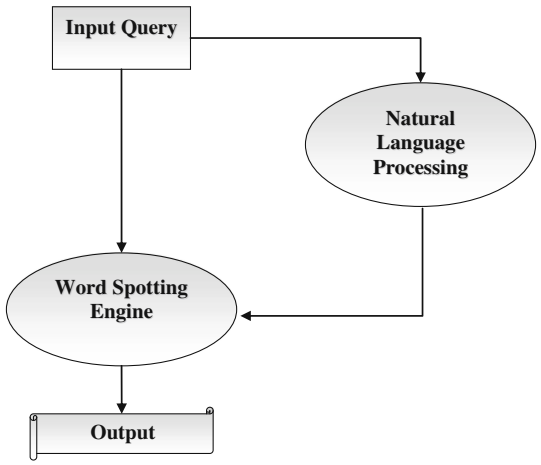

Fig. 1 System architecture

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 that provides users the ability to search documents using only a base word-form and locate all the corresponding inflected words.
ii a synonym dictionary that facilitates the access to the semantic context of documents and enriches the results of the search process.

The novelty of the proposed approach consists of:

- a word image retrieval system (word spotting) that links natural language processing techniques with document image analysis and search for the improvement in the retrieval accuracy
- the use of a combined word segmentation methodology based on two complementary segmentation techniques which enhances performance
- a generator of inflected word-forms that incorporates finite state rules aiming at covering the principals of the nominal inflection phenomenon of an early stage of Modern Greek which—to our knowledge—has not yet been systematically analyzed
- extensive evaluation of the proposed word spotting for early Greek historical documents.

3 Word spotting

In the case of historical documents, Manmatha and Croft [26] 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 [21] 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 toward 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 [14] and consists of five distinct steps: a preprocessing 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 [43]. For skew detection, a fast Hough transform approach is used that is based on the description of binary images using rectangular blocks [28].

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 [36]. 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 [12]. An example of applying image preprocessing step is given in Fig. 2.

3.2 Segmentation

In the proposed methodology, word segmentation is accomplished with the use of two complementary segmentation methodologies. The first is the run length smoothing algorithm (RLSA) [41] which is applied by using dynamic parameters that depend on the average character height as described in [21]. 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 (Fig. 3a). In the sequel, a connected component analysis is applied in order to extract the final word segmentation result (Fig. 3b).

The second segmentation methodology that we use is based on projection profiles. A similar approach has been used for the word segmentation of historical and degraded machine-printed documents in [2]. At a first step, we calculate the horizontal projection profile by summing the foreground pixels in every scan line. After a smoothing procedure, we calculate the local minima that define the

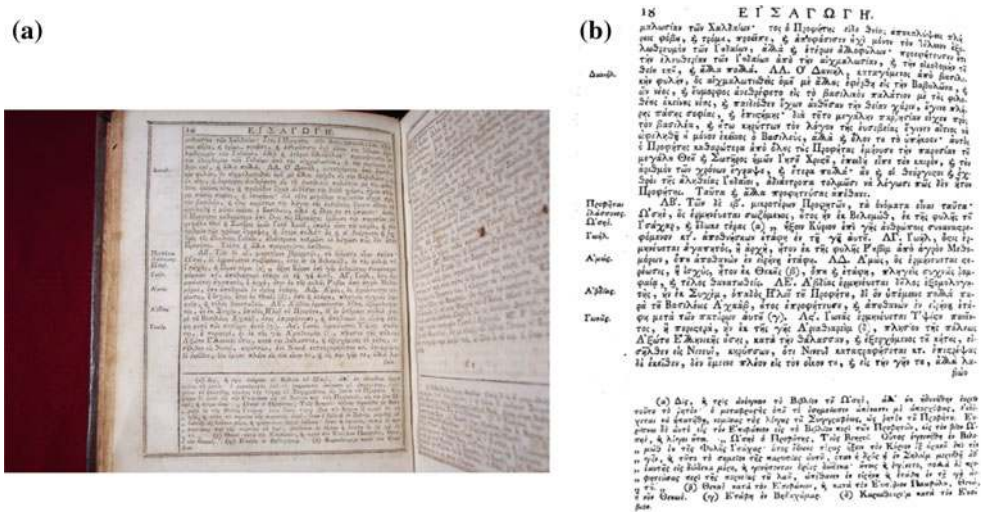

Fig. 2 **a** Original image; **b** image after preprocessing

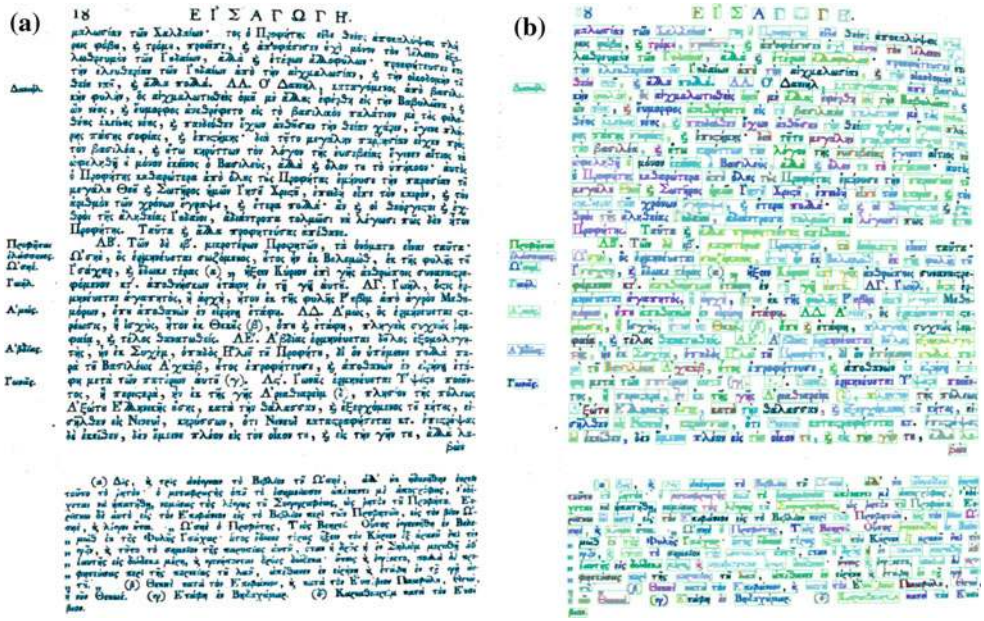

Fig. 3 Applying RLSA to the image of Fig. 2: **a** after smoothing; **b** word segmentation result

boundaries of the regions which contain the text lines (Fig. 4a). For every detected text line, similar procedure based on vertical projections is used in order to detect words and characters (Fig. 4b).

The combination of the two segmentation techniques is done as follows: Word matching is performed using word segmentation results from both methodologies. At the result list, if we have two results that correspond to different segmentation methodologies and their bounding boxes overlap, then we keep only the result that has the smaller distance to the query word (see Sect. 3.5).

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 equivalent. Thus, an image template for all required characters has to be previously defined and stored to the system. As shown in Fig. 5, 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

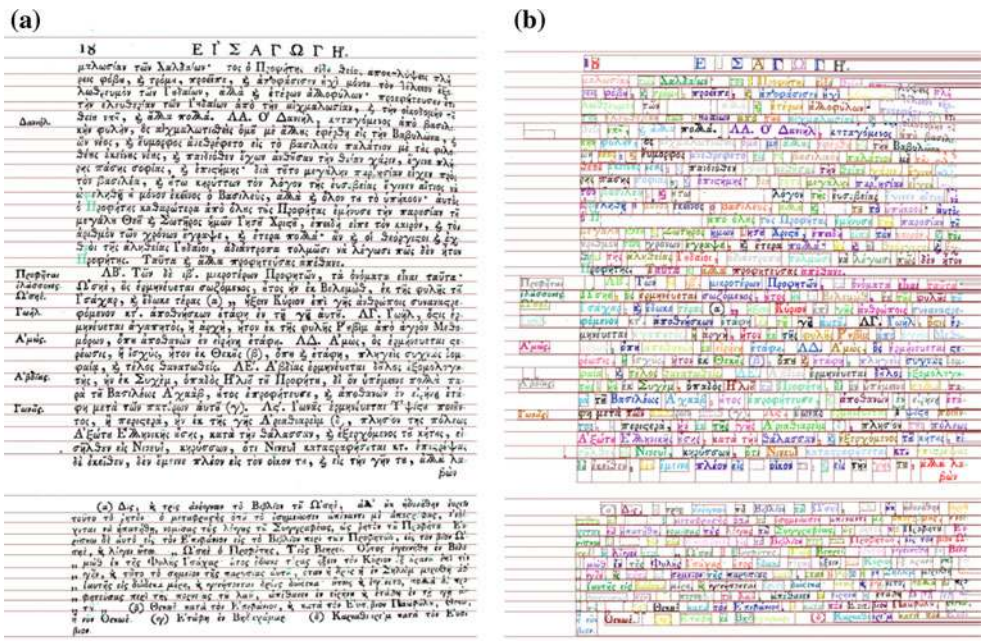

Fig. 4 Applying projection profiles application to the image of Fig. 2: **a** text line segmentation result; **b** word segmentation result

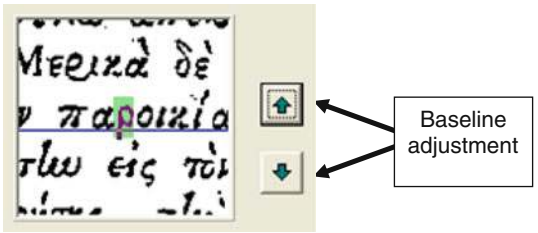

Fig. 5 Character template selection

to avoid inaccurate alignment results since there are characters whose lower end is below the text line, e.g. the letter "ρ" in Fig. 5. 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.

3.4 Feature extraction

The segmented words of the historical document as well as the synthetic query word are described by features that 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 [7,30]. In the proposed approach, two different types of features are used. The first one, which is based on [5], 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 that passes through the center of mass 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 [30]. The word is divided into 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

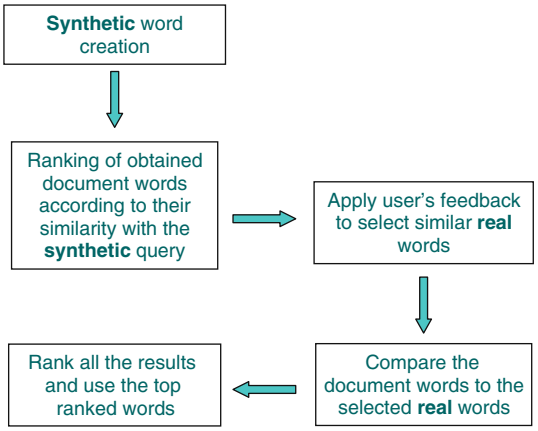

Fig. 6 The consecutive stages of the user feedback process

the indexed segmented words. Ranking of the comparison results is based on calculating the Euclidean distance metric using the above-mentioned features [37]. 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 6 illustrates the flow diagram of the user feedback process.

4 A morphological generator for early Modern Greek

Effective search in documents necessarily takes into account all possible inflected forms of a word without explicit formulation from the user. When search terms derive from a language with a rich morphological system like Greek, a morphological generation system based on a solid linguistic analysis is necessary.

The problem becomes even more complicated when dealing with historical documents where significant differences between the language of the text and the contemporary form of the language are observed. To this end, an attempt to deal with German historic texts with non-standard spelling is described in [8] where an algorithm for generating search term variants in ancient orthography of German is proposed.

In this paper, we describe an attempt to implement a generator of inflected word-forms, which incorporates finite state rules aiming at covering the principals of the nominal inflection phenomenon of an early stage of Modern Greek which—to our knowledge—has not yet been systematically analyzed. In this way, extensibility of the system is assured, and it can be easily enriched with new words simply by establishing the correspondence to the representative of each inflectional class. A preliminary implementation of a subset of these finite state rules is described in [22].

4.1 The linguistic approach

Our experimentation is performed on a corpus that consists of digitized historical and religious documents printed during the seventeenth and eighteenth 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. Although word structure in Modern Greek and earlier stages of the language remains the same, we came across certain significant differences especially in the number of inflected word-forms mainly due to two reasons:

- The language in the particular time period contains a richer set of inflectional endings since the number of inflectional affixes has diminished in the course of the evolution of the language.
- Certain grammatical cases appearing the text, such as the dative case, have disappeared in the current stage of the Greek language.

The set of keywords used in the experiment entails historical, theological, philological and linguistic interest. These words are mainly nouns of relevant high frequency of appearance which characterize the documents. Moreover, some grammatical elements of particular linguistic interest were taken into account. 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 above-mentioned 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 and information extraction [17]. Finite state approaches to morphology proved to be particularly successful covering a large spectrum of languages, including English, French, German, Dutch, Italian, Greek, Spanish, Portuguese, Finnish, Turkish, Arabic and other languages [1,16,17,20,31]. Morphological phenomena in Modern Greek have been extensively dealt with within the finite state framework, covering inflection, derivation and compounding [29,35]. 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.

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 [33,34]—a collection of software tools for the generation, manipulation and processing of finite-state transducers that 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 [4]. We opted for the SFST system because it is 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 [45]. It has been compiled using the MinGW and MSYS open software that can be downloaded from [46] and served as the basis of the SFST compiler.

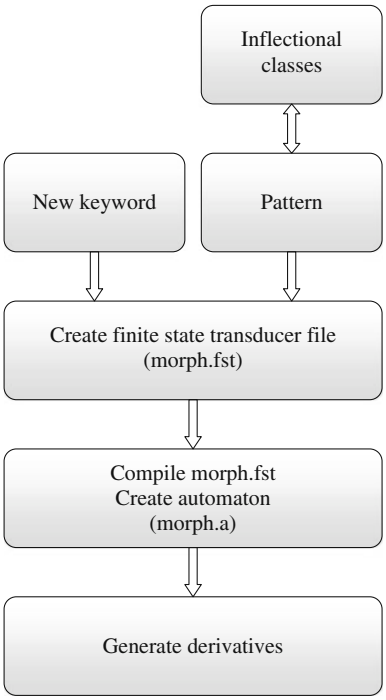

Fig. 7 Architecture of the morphological generator

4.3 The morphological generator

4.3.1 The system architecture

The architecture of the morphological generation tool is illustrated in Fig. 7. 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 finite state transducer file *Morph.fst* is created and dynamically compiled into the *morph.a* automaton that is responsible for the generation of word-forms.

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:

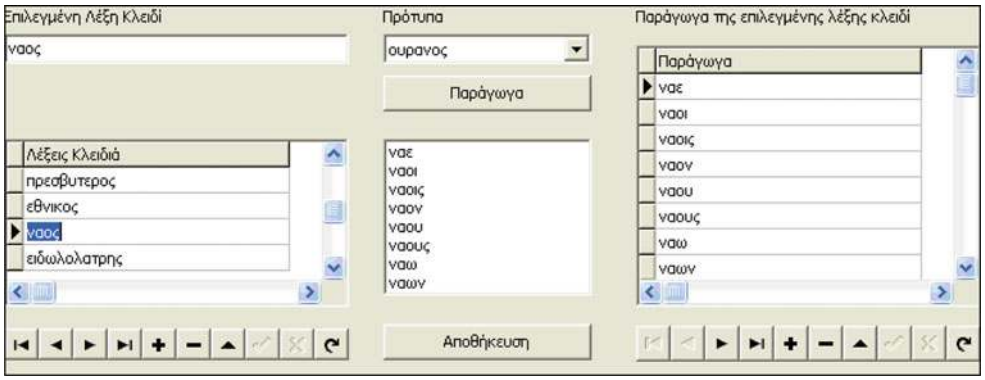

Fig. 8 Example of the morphological generation process. The keyword is "ναος", the inflectional class representative is "ουρανος" and the inflected word-forms are "ναε, ναοι, ναοις, ναον, ναου, ναους, ναω, ναων"

i. insert a new keyword
ii. assign the appropriate inflectional class through the selection of the appropriate class representative as illustrated in Fig. 8 where the word 'ναος' ('temple') is inserted, and the appropriate inflectional class is selected according to the pattern 'ουρανος' ('heaven / sky')
iii. perform a generation of all inflected word-forms according to the function of the selected representative of the inflectional class as illustrated in Fig. 8.

At this stage of the processing, the user need 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 that are used in the word spotting procedure.

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 in [40]. Yet, the usage of the particular semantic network has been often questioned as for the performance improvement in the query expansion procedure. Therefore, the selection of the terms in the query expansion is performed either from manually created thesauri such as WordNet [23,40,44] or from co-occurrence-based ones [3,10,11]. A hybrid approach taking into account both hand-crafted and co-occurrence-based thesauri is undertaken in [6] and [25]. As it has been recently proved in [9], exclusive use of manually created thesauri can lead to significant performance improvement.

In order to improve the efficiency of our system, we had to take into account the fact that the terms in the historical documents in the collection are drawn from the specific religious vocabulary. So, 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 [38].

Our approach is based on the methodology proposed in [38] in particular in ranking the synonymy relations in terms of their weight in the religious—historical domain in order to associate them with a score that estimates how often the relation is used in the specific domain. Relations that are irrelevant to the domain are not inserted in the synonymy database.

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. Figure 9 illustrates the synonym dictionary update procedure.
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 as illustrated in Fig. 10. The ensemble of words found (original word and the corresponding synonyms) can be used to update the query in the word spotting process.

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

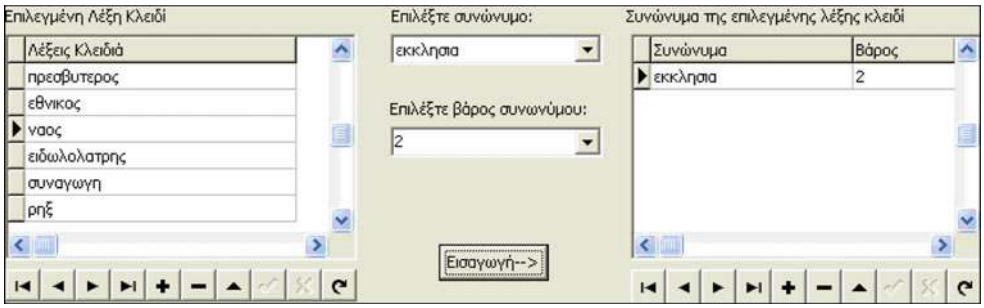

Fig. 9 Synonym dictionary update. The keyword is "$\nu\alpha o\varsigma$" ('temple'), and the synonym inserted is "$\varepsilon\kappa\kappa\lambda\eta\sigma\iota\alpha$" ('church') with a weighted value of 2

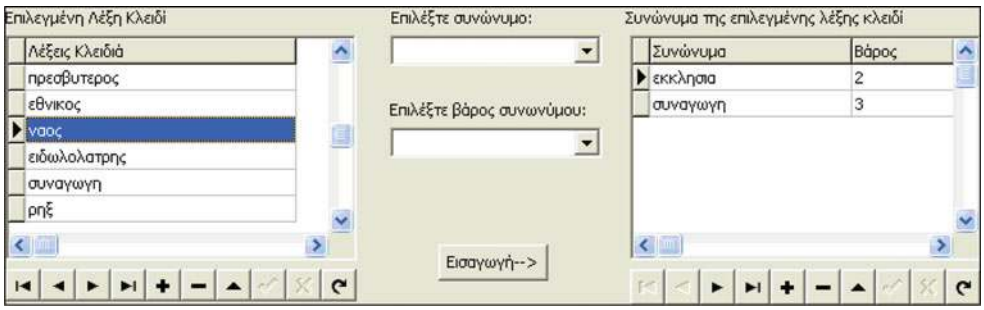

Fig. 10 Synonym dictionary lookup. The keyword is "$\nu\alpha o\varsigma$" ('temple'), and the synonyms are "$\varepsilon\kappa\kappa\lambda\eta\sigma\iota\alpha$" ('church') with a weighted value of 2 and "$\sigma\upsilon\nu\alpha\gamma\omega\gamma\eta$" ('synagogue') with a weighted value of 3

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.

6 Experimental results

The experiments addressed in this work are based on a corpus of 110 images which correspond to single pages of historical Greek documents printed during the seventeenth and eighteenth century. The query set consists of 32 keywords. For each keyword, the corresponding inflected word-forms have been determined using the proposed morphological generator. The keywords and the corresponding inflected word-forms that appear in the document corpus are shown in Table 1.

The overall ground truth has been determined in advance identifying all the instances of the keywords and the corresponding word-forms. Throughout the whole corpus (110 document pages), there exist 757 word instances of interest. The image preprocessing steps described in Sect. 3.1 are applied to each document image separately. The document images are segmented at word level using two techniques presented in Sect. 3.2, namely RLSA and projections profiles. The overall number of segmented words in each case is 48,912 and 57,881 words, respectively.

6.1 Word segmentation evaluation

This experiment concerns the evaluation of the word segmentation step which is critical for the complete word spotting procedure. The evaluation takes into account the segmentation techniques used, namely the RLSA and the projection profiles. First, the performance of each segmentation methodology is evaluated separately and then the evaluation takes into account the combination of both segmentation techniques. The word segmentation accuracy is calculated for all word instances that correspond to keywords and the corresponding inflected word-forms as defined in the ground truth. A word instance is considered as detected only if there is a significant overlap with an existing segmentation result. The overlap is expressed by the intersection over union metric IOU $= \frac{A \cap B}{A \cup B}$ where A and B denote the bounding box areas of the word instance and a segmented word, respectively [39,42]. The IOU metric ranges from 0 to 1, where 1 corresponds to exact matching. A threshold T is applied in order to decide whether the word instance and the segmented word match sufficiently. In the case where the segmentation methods are used in a combined way and a word instance matches sufficiently a segmented keyword from either

Table 1 Keywords along with the corresponding inflected word-forms and the overall number of actual instances in the document corpus

Keyword	Inflected word-forms	Instances
αμαρτία	αμαρτίας, αμαρτιών	6
απόστολος	αποστόλοις, αποστόλου, αποστόλους, αποστόλων, απόστολοι, απόστολον	52
άρχων	άρχοντα, άρχοντας, αρχόντων	5
αυτοκράτωρ	αυτοκράτορα, αυτοκράτορας, αυτοκράτορες, αυτοκράτορος, αυτοκρατόροις, αυτοκρατόρων	80
βάσανο	βάσανα, βασάνων	5
βασιλεύς	βασιλέα, βασιλέας, βασιλέων, βασιλέως, βασιλείς,	60
γη	γης, γην	26
εθνικός	εθνικοί, εθνικοίς, εθνικούς, εθνικόν, εθνικών	22
εκκλησία	εκκλησίαι, εκκλησίαν, εκκλησίας, εκκλησιών	138
έλληνας	έλληνες, ελλήνων	14
έργο	έργα, έργω, έργων	9
ευαγγελιστής	ευαγγελιστήν, ευαγγελισταί, ευαγγελιστού, ευαγγελιστών	9
θάλασσα	θάλασσαν, θαλάσσης	3
θάνατος	θάνατον, θανάτου, θανάτω	51
θυσία	θυσίας	7
κακία	κακίαν, κακίας	5
κανόνας	κανόνα, κανόνες, κανόνι, κανόνων	7
κληρικός	κληρικοί, κληρικούς, κληρικών	7
κολαστήριο	κολαστήρια, κολαστηρίων	6
λατίνος	λατίνοι, λατίνους, λατίνων	19
μαρτυρία	μαρτυρίου, μαρτυρίων, μαρτύριον	18
ναός	ναοί, ναού, ναούς, ναόν	15
νόμος	νόμον, νόμου, νόμους	24
πίστις	πίστει, πίστεως, πίστιν	36
πράξις	πράξει, πράξεις, πράξεων, πράξιν	17
πρεσβύτερος	πρεσβυτέροις, πρεσβυτέρου, πρεσβυτέρων, πρεσβύτεροι, πρεσβύτερον, πρεσβύτερους	24
ρηξ	ρήγας, ρήγαν	4
ρωμαίος	ρωμαίοι, ρωμαίοις, ρωμαίον, ρωμαίου, ρωμαίους, ρωμαίων	45
συναγωγή	συναγωγάς, συναγωγήν, συναγωγής	15
σφάλμα	σφάλματα	3
τόπος	τόπον, τόπους, τόπω, τόπων	18
χώρα	χώραν, χώρας	7

segmentation method, then the one with the highest IOU value is considered. Figure 11 shows the word segmentation results when applied to the whole document corpus using the segmentation methods separately as well as combined. Five different values of threshold values T have been applied ranging from 0.5 up to 0.9 by a step of 0.1. It is clearly shown in Fig. 11 that in all cases, the combination of the aforementioned segmentation techniques results in significantly improved word segmentation accuracy.

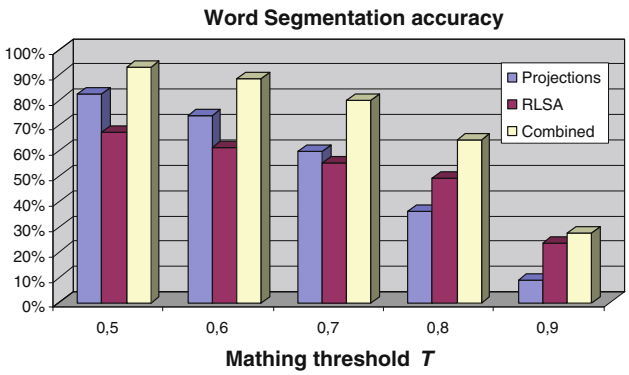

Fig. 11 Word segmentation evaluation

6.2 Word spotting

In the following experiment, the word spotting procedure is evaluated by searching instances of each keyword in the set of segmented words. For each keyword, a synthetic word image representation is created using character templates that correspond to the ASCII equivalent of each character. The synthetic word image features are matched against the features of each segmented word image in the document corpus, and finally a ranking is given based upon their similarity. This process is applied separately for each one of the two segmentation methods. In order to retrieve the best results, the two ranking lists are combined and a re-ranking is applied. In the course of this procedure, special consideration should be given, so that any multiple instances of the same segmented word are removed from the combined ranking list. The rule applied defines that only the word with the highest similarity to the synthetic keyword is retained.

The initial ranked list is used in order to select the correct words of the corpus according to the user's feedback

Fig. 12 **a** Synthetic keyword "Βασιλέως" **b** Retrieval results without user feedback; **c–e** Retrieval results when the top one, two and three relevant instances are selected by the user, respectively

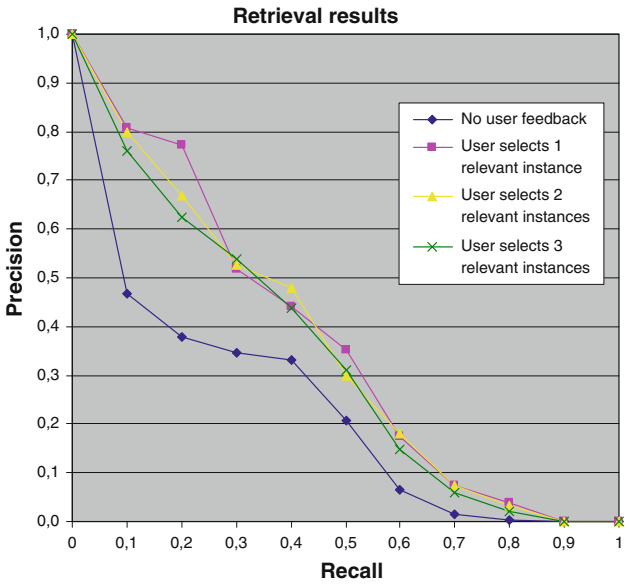

Fig. 13 Average retrieval results for all the query keywords

Table 2 Retrieval statistics

	No user feedback (%)	1 relevant instance (%)	2 relevant instances (%)	3 relevant instances (%)
1st Tier	36	39	42	41
2nd Tier	39	44	49	45
1-NN	67	83	83	83

performing a transition from synthetic to real data. In this case, the segmented words are ranked according to their similarity to the selected real word(s) of the document's corpus. In our experiments, the user's feedback was emulated by selecting the top one, two and three relevant instances from the initial ranked list. For each selected real word, a new matching process is initiated. The matching results for the one, two or three selected words are merged and re-ranked in order to get the final ranked list. In all cases, where user feedback is involved, the "freezing" method [19,32] is applied during the evaluation process. In the traditional freezing approach, any list entries identified as relevant by the user are kept "frozen" in their original ranks. The purpose of this process is that because of freezing, the relevant instances selected by the user are not re-retrieved with better ranks.

Figure 12 shows retrieval results for the query keyword "Βασιλέως". The synthetic keyword composed by character templates is shown in Fig. 12a. Figure 12b shows the first 15 entries in the ranking list without user feedback. It contains the segmented words ranked according to their similarity to the synthetic keyword. A check mark denotes segmented words that are relevant instances of the keyword according to the ground truth. The first three relevant words are marked by a red bounding box. The user's feedback effect is demonstrated in Fig. 12c–e. In particular, Fig. 12c shows the retrieval results when the user selects the first relevant instance as appear in Fig. 12b and an updated matching process is applied. Similarly, Fig. 12d–e show the retrieval results when the user selects the first two and three relevant instances, respectively.

Figure 13 shows the retrieval results taking into account all the keywords as average precision-recall diagrams. There are four curves that correspond in the case that there is no user feedback as well as in the case that user feedback is applied for one, two and three relevant instances, respectively. It can be observed that the user's feedback provides a significant boost to the accuracy of the retrieval results in terms of precision/recall. However, using more than one relevant instance may not further increase the retrieval performance. Indeed, due to the freezing method which is applied in the evaluation process, the relevant words that correspond to the user's selection are kept to their "frozen" positions instead of being moved to the top positions of the ranking list. This may affect the overall retrieval performance, especially for keywords that have few instances in the documents corpus. Table 2 provides typical retrieval performance measures, namely the 1st Tier, 2nd Tier and the 1-NN metric, in the case of the aforementioned user involvement.

6.3 Word spotting using word-forms

In the next set of experiments, the proposed methodology is tested by taking into account the inflected word-forms of each query keyword. For this experimentation, the user's

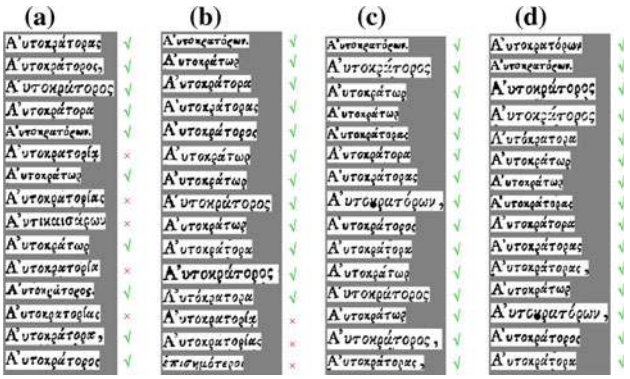

Fig. 14 Retrieval results for keyword "Αυτοκράτωρ" using inflected word-forms **a** Retrieval results without user feedback; **b–d** Retrieval results when the top one, two and three relevant instances are selected by the user, respectively

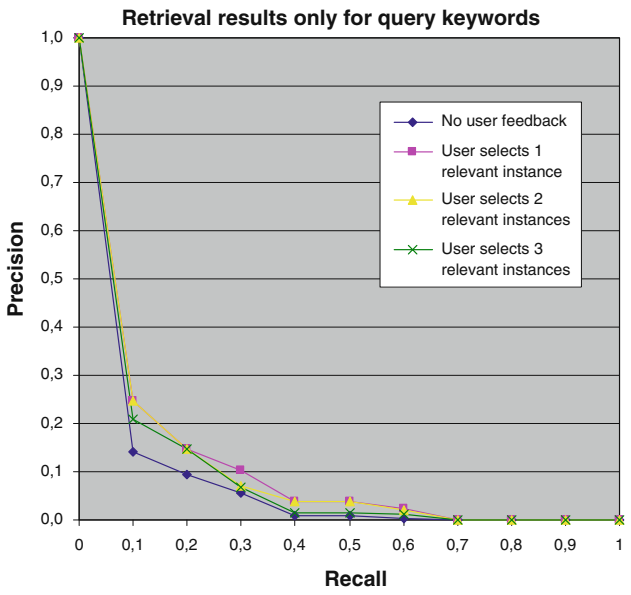

Fig. 15 Average retrieval results for all the keywords when no inflected word-forms are considered

Table 3 Retrieval statistics without inflected word-forms

	No user feedback (%)	1 relevant instance (%)	2 relevant instances (%)	3 relevant instances (%)
1st Tier	11	14	14	12
2nd Tier	12	15	15	15
1-NN	21	37	32	26

feedback process is not only applied to the keyword but it is also applied to its inflected word-forms. For each of the four working cases (i.e. no feedback, one, two, or three selected relevant words), the final ranking results of the keyword and its inflected word-forms are merged and re-ranked. In this case, the top entries of the final ranking list are instances of the keyword itself or its inflected word-forms.

Figure 14 shows retrieval results for the query keyword "Αυτοκράτωρ". According to Table 1, this keyword has 6 inflected word-forms. Figure 14a shows the first 15 entries of the initial ranking results without user feedback. The check mark denotes segmented words that are relevant instances of the keyword or of one of its inflected word-forms. The three columns in Fig. 14b–d show the retrieval results when the user selects the first one, two or three relevant instances, respectively. Again, the user's feedback increases the retrieval performance.

To provide a comparative evaluation of the word spotting procedure using inflected word-forms, we aim in addressing retrieval performance in two distinct cases that should be compared against each other. In the first case, a retrieval process is performed in which instances of each query keyword are used without taking into consideration any inflected word-forms, while in the second case, a retrieval process is performed in which it is not only instances of each query keyword which are used but also any corresponding inflected word-forms are taken into consideration. For both cases, the retrieval results are evaluated with respect to the ability of the particular retrieval process to return all relevant word images to the query keyword along with the corresponding inflected word-forms as defined in the ground truth.

Figure 15 shows the retrieval results for all the query keywords in terms of average precision/recall when no inflected word-forms are used (first case). The four curves correspond to no user feedback, one, two and three selected relevant instances, respectively. Table 3 provides further retrieval statistics for this step. It is clearly shown that for this case a low performance is achieved. Figure 16 shows the retrieval results for all the query keywords in terms of average precision/recall when inflected word-forms are used (second case). Additional retrieval statistics for this particular case are given in Table 4. It can be observed that using inflected word-forms in the word spotting procedure results in significantly improved retrieval performance when compared to the case where only the query keywords are considered. Table 5 shows processing time requirements. The method has been implemented in Matlab®, and the vast majority of computational time concerns the comparison of the synthetic or relevant keyword against the 1,06,793 segmented words from all the pages of the document corpus. It should be noticed that these time expenses do not include the time slot that concerns the off-line process, namely the word segmentation and the feature extraction for all the segmented words in the database.

7 Conclusions

In this paper, we have described a methodology for accessing the content of digitized Greek historical documents without

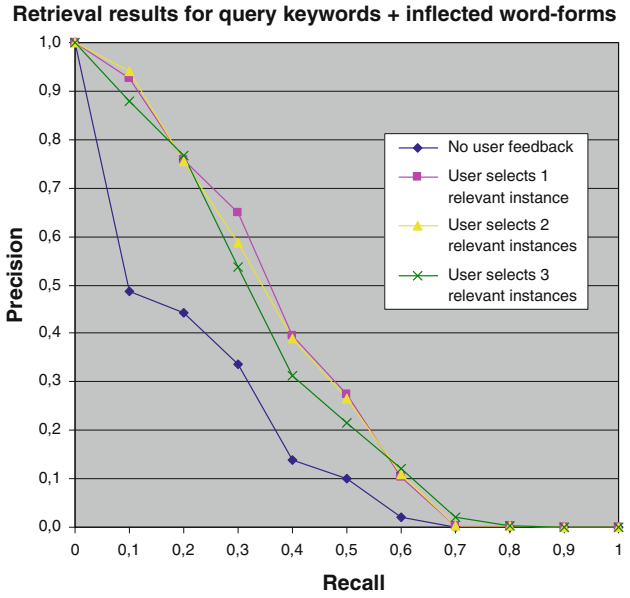

Fig. 16 Average retrieval results for all the keywords when inflected word-forms are also considered

Table 4 Retrieval statistics with inflected word-forms

	No user feedback (%)	1 relevant instance (%)	2 relevant instances (%)	3 relevant instances (%)
1st Tier	25	40	38	36
2nd Tier	29	45	44	44
1-NN	50	88	88	75

Table 5 Computational requirements per keyword

	Computational time (s)
Synthetic keyword Matching + Ranking	18
Three relevant instances Matching + Ranking	41

the use of an optical character recognition system. Access to the documents that were printed during the seventeenth and eighteenth century is based on word spotting using word images created from keywords.

At a first step, we apply a preprocessing step in order to improve the quality of the document image, while word segmentation is accomplished with the use of two complementary segmentation methodologies. Then, 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. The experimental results show that using inflected word-forms in the word spotting procedure along with the user's feedback leads to improved retrieval performance when compared to the case where only the query keywords are considered.

Acknowledgments The research leading to these results has received funding from the Greek Ministry of Research funded R&D (POLYTIMO project) as well as from the European Community's Seventh Framework Programme under grant agreement No. 215064 (IMPACT project).

References

1. Antworth, E.: PC-KIMMO: A Two-level Processor for Morphological Analysis, Occasional Publications in Academic Computing no 16, Summer Institute of Linguistics, Dallas TX (1990)
2. Antonacopoulos, A., Karatzas, D.: Semantics-based content extraction in typewritten historical documents. In: Eighth International Conference on Document Analysis and Recognition, pp. 48–53, 2005
3. Bai, D., Song, P., Bruza, J., Nie, J., Cao, J.: Query expansion using term relationships in language models for information retrieval. In: Proceedings of the 14th International Conference on Information and Knowledge Management (CIKM05), 2005
4. Beesley, K., Karttunen, L.: Finite State Morphology. CSLI Publications, Stanford (2003)
5. Bokser, M.: Omnidocument technologies. Proc. IEEE **80**(7), 1066–1078 (1992)
6. Cao, J., Nie, J., Bai, J.: Integrating word relationships into language models. In: Proceedings of the 2005 ACM SIGIR Conference on Research and Development in Information Retrieval, 2005
7. Doerman, D.: The detection of duplicates in document image databases. In: Proc. of the 4th Int. Conf. on Document Analysis and Recognition (ICDAR'97), pp. 314–318, 1997
8. Ernst-Gerlach, A., Fuhr, N.: Generating Term Variants for Text Collections with Historic Spellings. In: Proceedings of the 28th European Conference on Information RetrievalResearch (ECIR 2006), Springer, 2006
9. Fang, H.: A re-examination of query expansion using lexical resources. In: Proceedings of ACL'08, pp. 139–147, Columbus, Ohio, 2008
10. Fang, H., Zai, C.: An exploration of axiomatic approaches to information retrieval. In: Proceedings of the 2005 ACM SIGIR Conference on Research and Development in Information Retrieval, 2005
11. Fang, C., Zai, C.: Semantic term matching in axiomatic approaches to information retrieval. In: Proceedings of the 2006 ACM SIGIR Conference on Research and Development in Information Retrieval, 2006
12. Gatos, B., Danatsas, D., Pratikakis I., Perantonis, S.J.: Automatic table detection in document images. In: Proceedings of the

Third International Conference on Advances in Pattern Recognition (ICAPR'05). Lecture Notes in Computer Science (3686), pp. 609–618. (2005)
13. Gatos, B., Papamarkos, N., Chamzas, C.: A binary tree based OCR technique for machine printed characters. Eng. Appl. Artif. Intell. **10**(4), 403–412 (1997)
14. Gatos, B., Pratikakis, I., Perantonis, S.J.: Adaptive degraded document image binarization. Pattern Recognit **39**, 317–327 (2006)
15. Guillevic, D., Suen, C.Y.: HMM word recognition engine. In: Fourth International Conference on Document Analysis and Recognition (ICDAR'97), pp. 544–547, 1997
16. Karttunen, L.: KIMMO: a general morphological processor. Tex. Linguist. Forum **22**, 163–186 (1983)
17. Karttunen, L., Oflazer, K.: Special issue on finite-state methods in NLP: computational linguistics. **26**(1), 1–2 (2000)
18. Keaton, P., Greenspan, H., Goodman, R.: Keyword spotting for cursive document retrieval. In: Workshop on Document Image Analysis (DIA 1997), pp. 74–82, 1997
19. Keskustalo, H., Järvelin, K., Pirkola, A.: Evaluating the effectiveness of relevance feedback based on a user simulation model: effects of a user scenario on cumulated gain value. Inf. Retr. **11**(3), 209–228 (2008)
20. Koskenniemi, K.: Two-level Morphology: A General Computational Model for Word-form Recognition and Production. Publication No 11, Dept. of General Linguistics, University of Helsinki (1983)
21. Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I., Theodoridis, S., Perantonis, S.J.: Keyword-guided word spotting in historical printed documents using synthetic data and user feedback. Int. J. Doc. Anal. Recognit. (IJDAR) Spec. Issue Hist. Doc. **9**(2–4), 167–177 (2007)
22. Lampropoulos, A., Galiotou, E., Manolessou, I., Ralli, A.: A finite state approach to the computational morphology of early Modern Greek. In: Proceedings of the 7th WSEAS International Conference on Applied Computer Science, Venice, pp. 242–245, 2007
23. Liu, S., Liu, F., Yu, C., Meng, W.: An effective approach to document retrieval using WordNet and recognizing phrases. In: Proceedings of the 2004 ACM SIGIR Conference on Research and Development in Information Retrieval, 2004
24. Lu, Y., Tan, C., Weihua, H., Fan, L.: An approach to word image matching based on weighted Hausdorff distance. In: Sixth International Conference on Document Analysis and Recognition (ICDAR'01), pp. 10–13, 2001
25. Mandala, R., Tokunaga, T., Tanaka, H.: Combining multiple evidence from different types of thesaurus for query expansion. In: Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval, 1999
26. Manmatha, R., Croft, W.B.: A Draft of Word Spotting: Indexing Handwritten Manuscripts, Intelligent Multimedia Information Retrieval. pp. 43–64. MIT Press, Cambridge, MA (1997)
27. Marcolino, A., Ramos, V., Ármalo, M., Pinto, J.C.: Linea and Word matching in old documents. In: Proceedings of the Fifth Ibero-American Symposium on Pattern Recognition (SIAPR'00), pp. 123–125, 2000
28. Perantonis, S.J., Gatos, B., Papamarkos, N.: Block decomposition and segmentation for fast Hough transform evaluation. Pattern Recognit. **32**(5), 811–824 (1999)
29. Ralli, A., Galiotou, E.: Greek Compounds: A Challenging Case for the Parsing Techniques of PC-KIMMO v.2. Int. J. Comput. Intell. **1**(2), 152–162 (2004)
30. Rath, T.M., Manmatha, R.: Features for word spotting in historical documents. In: Proc. of the 7th Int. Conf. on Document Analysis and Recognition (ICDAR'03), pp. 218–222, 2003
31. Roark, B., Sproat, R.: Computational Approaches to Morphology and Syntax. Oxford university Press, Oxford (2007)
32. Salton, G.: Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading (1989)
33. Schmid, H.: A Programming Language for Finite State Transducers. In: Proc. FSMNLP 2005, Helsinki, Finland, 2005
34. Schmid, H., Fitschen, A., Heid, U.: SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In: Proc. LREC 2004, Lisbon, Portugal, pp. 1263–1266, 2004
35. Sgarbas, K., Kokkinakis, N.G.: A PC-KIMMO-Based Morphological Description of Modern Greek. Lit. Linguist. Comput. **10**(3), 189–201 (1995)
36. Stamatopoulos, N., Gatos, B., Kesidis, A.: Automatic Borders Detection of Camera Document Images. In: 2nd International Workshop on Camera-Based Document Analysis and Recognition (CBDAR'07), Curitiba, Brazil, pp. 71–78, 2007
37. Theodoridis, S., Koutroumbas, K.: Pattern recognition. Academic Press, New York (1997)
38. Turcato, D., Popowich, F., Toole, J., Fass, D., Nicholson, D., Tisher, D.: Adapting a synonym database to specific domains. In: Klavans J., Gonzalo J. (eds.) Proceedings of the ACL Workshop on Recent Advances in Natural Language Processing and Information Retrieval, pp. 1–11 (2000)
39. Veltkamp, R.C., Hagedoorn, M.: Shape similarity measures, properties, and constructions. In: Advances in Visual Information Systems, 4th Int. Conf, VISUAL 2000, pp. 467–476, 2000
40. Voorhees, E.M. : Using WordNet for text retrieval. In: Fellbaum, C. (ed.) WordNet: An Electronic Lexical Database, chap. 12, pp. 285–303. MIT Press Books, Cambridge (1998)
41. Wahl, F.M., Wong, K.Y., Casey, R.G.: Block segmentation and text extraction in mixed text/image documents. Comput. Graph. Image Process **20**, 375–390 (1982)
42. Wolf, C., Jolion, J.: Object count/area graphs for the evaluation of object detection and segmentation algorithms. Int. J. Doc. Anal. Recognit. **8**(4), 280–296 (2006)
43. Yin, P.Y.: Skew detection and block classification of printed documents. Image Vis. Comput. **19**, 567–579 (2001)
44. Zhiguo, G., Chan, W.C., Long, H.U.: Web query expansion by WordNet. In: Proceedings of DEXA'05, Copenhagen, pp. 166–175, Springer, 2005
45. ftp://ftp.ims.uni-stuttgart.de/pub/corpora/SFST/
46. http://www.mingw.org/