The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0024-2535.htm

Received 12 November 2013
Revised 13 January 2014
Accepted 26 January 2014

Using digital corpora for preserving and processing cultural heritage texts: a case study

Eleni Galiotou
Department of Informatics, Technological Educational Institute of Athens, Athens, Greece

Abstract

Purpose – The purpose of this paper is to describe the creation and exploitation of a historical corpus in an attempt to contribute to the preservation and availability of cultural heritage documents.

Design/methodology/approach – At first, the digitization process and attempts to the availability and awareness of the books and manuscripts in a historical library in Greece are presented. Then, processing and exploitation, taking into account natural language processing techniques of the digitized corpus, are discussed.

Findings – In the course of the project, methods that take into account the state of the documents and the particularities of the Greek language were developed.

Practical implications – In its present state, the use of the corpus facilitates the work of theologians, historians, philologists, paleographers, etc. and in the same time, prevents the original documents from further damage.

Originality/value – The results of this undertaking can give useful insights as for the creation of corpora of cultural heritage documents and as for the methods for the processing and exploitation of the digitized documents which take into account the language in which the documents are written.

Keywords Information retrieval, Preservation, Digitization, Natural language processing, Cultural heritage corpora, Greek language, Manuscripts

Paper type Case study

1. Introduction

Collections of manuscripts and early printed books of great philological and historical interest are often kept in libraries of monasteries and other remote historical sites. Access to such collections is quite a difficult task due to factors such as the deteriorated state of the collection items, the conditions under which the collections are kept and the limited accessibility to the location.

In recent years, the digitisation of such collections and their availability to the public has greatly facilitated research activities in the humanities. Historical corpora consisting of manuscripts and early printed books introduce various issues relating to their form and their language. Methods for dealing with these issues have been proposed

An earlier version of this paper was presented at the 2nd International Conference on Integrated Information, IC-ININFO, held in Budapest, Hungary, from 30 August to 2 September, 2012, http://history.icininfo.net/2012/

Emerald

Library Review
Vol. 63 No. 6/7, 2014
pp. 408-421
© Emerald Group Publishing Limited
0024-2535
DOI 10.1108/LR-11-2013-0142

in the course of research projects such as IMPACT, which aimed at massively digitising European historical texts and proposing methods for improving the access to their content (Balk and Conteh, 2011). Significant work has also been conducted on the digitisation and availability of Greek books and manuscripts.

For example, since 1985, the "Perseus" Digital Library[1] contains, among others, primary and secondary sources for the study of ancient Greece; in the course of the "Hellinomnimon"[2] project, the digital library of Greek philosophical and scientific books and manuscripts published from 1600 to 1821 was created in 2002; since 2006, the "Anemi" Digital Library[3] has provided access to a rich collection of digitised material relating to Modern Greek Studies. More recently, an attempt to create a digital library to accommodate archival and museum collections, located in the "Neophytos Doukas" municipal library, is reported in Varveris and Giannakopoulos (2009). The library in question is located in the village of Ano Pedina in the Zagori region of Northern Greece and contains a small but valuable historical collection. Therefore, it was used in this pilot project, aiming to achieve reorganisation while promoting and increasing the usage and services of a "typical" small village library.

In this paper, we present the historical corpus resulting from the process of digitisation of the texts in the library of the historical Holy Monastery of the Annunciation of the Virgin (in Greek *Ιεράά Μονή Ευαγγελισμού της Θεοτόκου - "Ευαγγελστρια"*) on the island of Skiathos, Greece. The collection consists of books printed during the 16th-19th centuries and manuscripts of the 13th to the 19th centuries. The thematic and lingual diversity of the collection items, as well as their bad state, give rise to a number of issues relating to the constitution of the digitised corpus and for its exploitation.

The paper is organised as follows: in Section 2, we give a brief description of the monastery and its library, and in Section 3, we describe the characteristics of the corpus and discuss issues regarding the availability and awareness of the digitised texts. Section 4 deals with the processing and exploitation of the corpus. In Section 5, we describe a set of Natural Language Processing tools for older varieties of Greek which contribute to the improvement of the search process. Finally, conclusions are drawn in Section 6.

2. The library

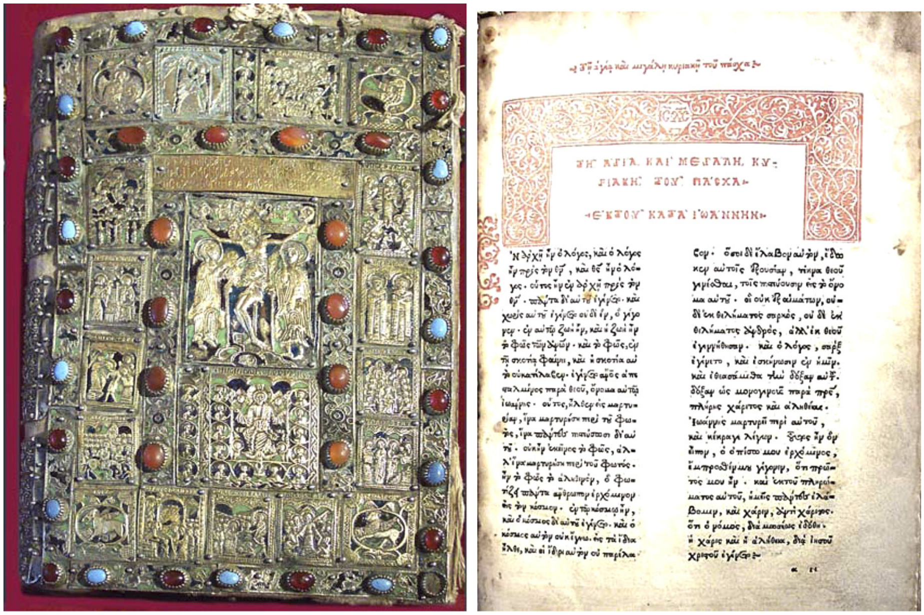
2.1 The monastery

The Holy Monastery of the Annunciation of the Virgin was founded on the island of Skiathos, Greece in 1794 by a group of monks, members of the "Kollyvades" spiritual movement, who were forced to leave the Holy Mount of Athos in the face of disturbances resulting from disputes that were taking place there. The "Kollyvades" movement had as an objective a return to the traditions of the ancient Orthodox Church and greatly influenced the spiritual life of the island. In addition, the monastery provided significant moral and material assistance to the Greek revolution of 1821 as well as to the pre-revolutionary movements. The cultural sites of the monastery include the religious and national museum, the folk museum and the exhibition which hosts Byzantine and post-Byzantine icons, wooden and silver crosses, liturgical vessels and other ecclesiastical items.

2.2 The content of the library

Of particular interest is the library where rare manuscripts and early printed books are kept. The library was formed by learned and bibliophile monks and other scholars who donated their books and manuscripts to the monastery. As a result, the collection is characterised by diversity as regards topics, dates, origin, etc. It comprises books printed from the 16th to 19th centuries, manuscripts from the 13th to 19th centuries and patriarchal documents. Obviously, most items in the collection are religious ones. For example, there are works by John Chrysostom, Basil the Great, Athanasius Parios, Nikodemos the Hagiorite and Patriarch Photius. Besides religious texts, there are also a significant number of books on topics of general interest such as Geography, Mathematics and Grammar, and annotated editions of works by Homer, Thucydides, Aristotle, etc. The oldest printed book is a hand-bound edition of the Gospel with gold, silver and precious stones, which was printed in Venice in 1539 (Plate 1).

Because of a long – but fortunately temporary – period of desolation of the monastery, the conditions in the library have drastically worsened. Therefore, the age of the items in the collection – especially the manuscripts – their deteriorated state and the conditions under which the items were kept created an urgent need to take action towards the preservation and awareness of the collection. The digitisation of the collection would enable access to the content of the collection items without further deterioration of their state before even proceeding to their preservation, which would ensure their survival.



Τῇ ἁγίᾳ καὶ μεγάλῃ κυριακῇ τοῦ πάσχα

ΤΗ ΑΓΙΑ ΚΑΙ ΜΕΓΑΛΗ ΚΥΡΙΑΚΗ ΤΟΥ ΠΑΣΧΑ.

ΕΚ ΤΟΥ ΚΑΤΑ ΙΩΑΝΝΗΝ.

Ἐν ἀρχῇ ἦν ὁ λόγος, καὶ ὁ λόγος ἦν πρὸς τὸν θεόν, καὶ θεὸς ἦν ὁ λόγος. οὗτος ἦν ἐν ἀρχῇ πρὸς τὸν θεόν. πάντα δι' αὐτοῦ ἐγένετο. καὶ χωρὶς αὐτοῦ ἐγένετο οὐδὲ ἕν, ὃ γέγονεν. ἐν αὐτῷ ζωὴ ἦν, καὶ ἡ ζωὴ ἦν τὸ φῶς τῶν ἀνθρώπων. καὶ τὸ φῶς, ἐν τῇ σκοτίᾳ φαίνει, καὶ ἡ σκοτία αὐτὸ οὐ κατέλαβεν. ἐγένετο ἄνθρωπος ἀπεσταλμένος παρὰ θεοῦ, ὄνομα αὐτῷ ἰωάννης. οὗτος ἦλθεν εἰς μαρτυρίαν, ἵνα μαρτυρήσῃ περὶ τοῦ φωτός, ἵνα πάντες πιστεύσωσι δι' αὐτοῦ. οὐκ ἦν ἐκεῖνος τὸ φῶς, ἀλλ' ἵνα μαρτυρήσῃ περὶ τοῦ φωτός. ἦν τὸ φῶς τὸ ἀληθινόν, ὃ φωτίζει πάντα ἄνθρωπον ἐρχόμενον εἰς τὸν κόσμον. ἐν τῷ κόσμῳ ἦν, καὶ ὁ κόσμος δι' αὐτοῦ ἐγένετο. καὶ ὁ κόσμος αὐτὸν οὐκ ἔγνω. εἰς τὰ ἴδια ἦλθε, καὶ οἱ ἴδιοι αὐτὸν οὐ παρέλαβον. ὅσοι δὲ ἔλαβον αὐτόν, ἔδωκεν αὐτοῖς ἐξουσίαν, τέκνα θεοῦ γενέσθαι, τοῖς πιστεύουσιν εἰς τὸ ὄνομα αὐτοῦ. οἳ οὐκ ἐξ αἱμάτων, οὐδὲ ἐκ θελήματος σαρκός, οὐδὲ ἐκ θελήματος ἀνδρός, ἀλλ' ἐκ θεοῦ ἐγεννήθησαν. καὶ ὁ λόγος, σὰρξ ἐγένετο, καὶ ἐσκήνωσεν ἐν ἡμῖν. καὶ ἐθεασάμεθα τὴν δόξαν αὐτοῦ. δόξαν ὡς μονογενοῦς παρὰ πατρός, πλήρης χάριτος καὶ ἀληθείας. ἰωάννης μαρτυρεῖ περὶ αὐτοῦ, καὶ κέκραγε λέγων. οὗτος ἦν ὃν εἶπον, ὁ ὀπίσω μου ἐρχόμενος, ἔμπροσθέν μου γέγονεν, ὅτι πρῶτός μου ἦν. καὶ ἐκ τοῦ πληρώματος αὐτοῦ, ἡμεῖς πάντες ἐλάβομεν, καὶ χάριν, ἀντὶ χάριτος. ὅτι ὁ νόμος, διὰ μωσέως ἐδόθη. ἡ χάρις καὶ ἡ ἀλήθεια, διὰ ἰησοῦ χριστοῦ ἐγένετο.

α ιι

Plate 1.
The Gospel (Venice, 1536)

Source: Library of the Holy Monastery of the Annunciation of the Virgin, Skiathos, Greece. Photo taken by author

2.3 Digitisation

Recently, the task of digitizing the library was undertaken in the course of the effort towards the maintenance and reconstruction of the monastery. The main reason for this undertaking was the increasing interest of philologists, theologians, historians and paleographers in the books and manuscripts of the collection and, at the same time, the very bad state of the items. Therefore, our goal was twofold:

- to prevent any further damage of the items; and
- to enable access to the texts and facilitate the work of researchers.

At first, an initial classification of the texts was performed, according to the "Catalog of Manuscripts and Books of the Holy Monastery of the Annunciation of the Virgin in Skiathos" which was created in 1916 by Anthony, former Bishop of Elia, and published after his death (Anthony, 1961, 1962a, 1962b). Although the catalogue was not created by a librarian, it constitutes the only systematic inventory of the library contents. Therefore, it was used to locate, classify and arrange the collection items on the library shelves. An appropriate database was created which would contain all the elements of the catalogue and indicate the location of the items on the shelves. Each record of the database would contain the following fields: serial number, title, author/editor, language, publisher, publication year, publication country, volume number, number of pages, comments and library/shelf number.

Following this, the texts were digitised and linked to the database. The extremely bad state and the fragility of the collection items have led us to avoid the use of a conventional scanner, whereas the lack of funding has prevented us from acquiring more sophisticated equipment. Therefore, the digitisation was performed using just a digital camera and a tripod. An important outcome of the digitisation procedure was the awareness of rare texts and of their extremely bad condition due to damage, stains, bleed-through distortion, annotations by readers, etc.

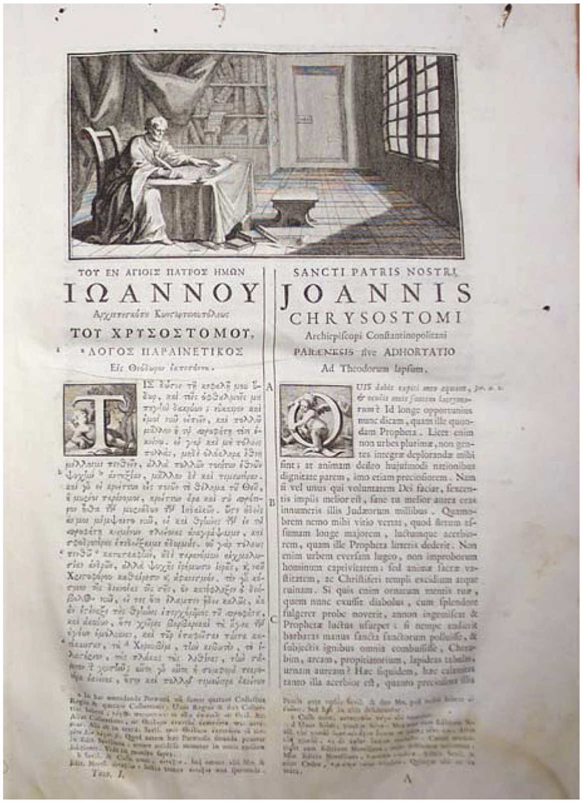
In the catalogue, 480 printed books (of which 144 were characterised as similar to other books in the collection) and 110 manuscripts are reported. Unfortunately, some 15 manuscripts and 50 printed books, which were reported in the catalogue, were lost. On the other hand, some eight manuscripts that were located in the library were not reported in the catalogue. Ultimately, 114 manuscripts and 289 printed books were digitised, resulting in about 189,000 jpeg images corresponding to single pages. Under the circumstances, it was not possible to adopt good practices in digitisation. Nevertheless, we plan to adopt such strategies, as they are reported in Frey and Lopes (2011) and Sotošek (2011), especially in generating metadata, in parallel with the process of preservation of the original collection items which will start in the immediate future.

3. The corpus

3.1 Corpus characteristics

As was mentioned in Section 2, the digitised collection of books and manuscripts is characterised by a thematic and linguistic diversity (Ancient, Medieval and Modern Greek). Other characteristics of the corpus can be summarised as follows:

- Most texts are monolingual (Greek). There are also a significant number of bilingual texts (Greek-Latin) (see also Plate 2). These bilingual texts have given insights regarding the development of tools for the alignment of Greek–Latin parallel texts (Sotiropoulos *et al.*, 2007).



ΤΟΥ ΕΝ ΑΓΙΟΙΣ ΠΑΤΡΟΣ ΗΜΩΝ
ΙΩΑΝΝΟΥ
Ἀρχιεπισκόπου Κωνσταντινουπόλεως
ΤΟΥ ΧΡΥΣΟΣΤΟΜΟΥ.
ΛΟΓΟΣ ΠΑΡΑΙΝΕΤΙΚΟΣ
Εἰς Θεόδωρον ἐκπεσόντα.

A Τίς δώσει τῇ κεφαλῇ μου ὕδωρ, καὶ τοῖς ὀφθαλμοῖς μου πηγὴν δακρύων [illegible] B [illegible] C [illegible]

[illegible]

Tom. I.

SANCTI PATRIS NOSTRI
JOANNIS
CHRYSOSTOMI
Archiepiſcopi Conſtantinopolitani
PARÆNESIS ſive ADHORTATIO
Ad Theodorum lapſum.

Quis dabit capiti meo aquam, & oculis meis fontem lacrymarum? Jer. 9. 1. Id longe opportunius nunc dicam, quam ille quondam Propheta. Licet enim non urbes plurimæ, non gentes integræ deplorandæ mihi ſint; at animam defleo hujuſmodi nationibus dignitate parem, imo etiam pretioſiorem. Nam ſi vel unus qui voluntatem Dei faciat, ſexcentis impiis melior eſt, B ſane tu melior antea eras innumeris illis Judæorum millibus. Quamobrem nemo mihi vitio vertat, quod fletum effuſurum longe majorem, luctumque acerbiorem, quam ille Propheta literis dederit. Non enim urbem everſam lugeo, non improborum hominum captivitatem: ſed animæ ſacræ vaſtitatem, ac Chriſtiferi templi excidium atque ruinam. Si quis enim ornatum mentis tuæ, quem nunc excuſſit diabolus, cum ſplendore C fulgeret probe noverit, annon ingemiſcat & Prophetæ luctus uſurpet; ſi nempe audierit barbaras manus ſancta ſanctorum polluiſſe, & ſubjectis ignibus omnia combuſſiſſe, Cherubim, arcam, propitiatorium, lapideas tabulas, urnam auream? Hæc ſiquidem, hæc calamitas tanto illa acerbior eſt, quanto pretioſiores illis

[illegible]

A

Plate 2.
Bilingual edition of John Chrysostom's works (Venice, 1734)

Source: Library of the Holy Monastery of the Annunciation of the Virgin, Skiathos, Greece. Photo taken by author

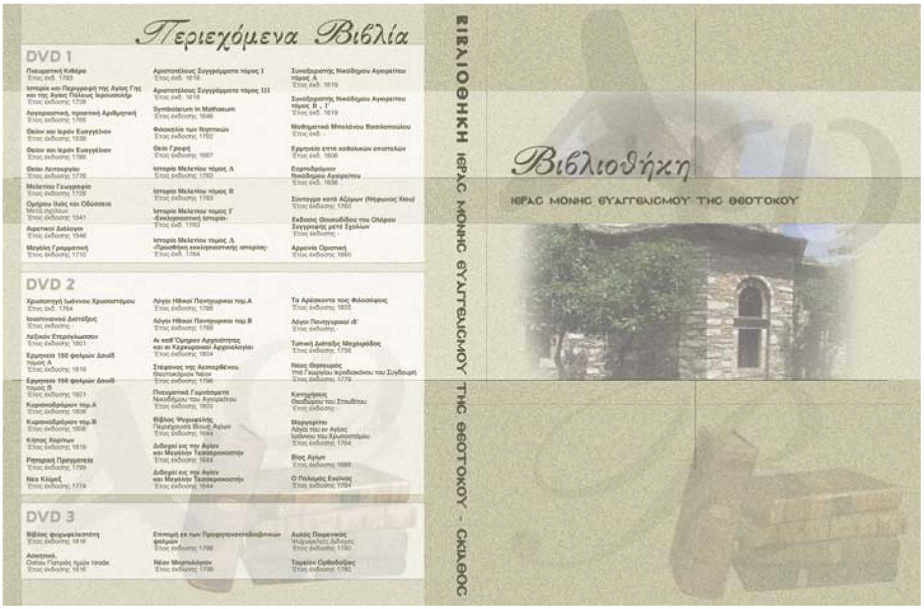
- The largest sub-corpus consists of religious texts and contains works by John Chrysostom, Basil the Great, Athanasius Parios, Nikodemos the Hagiorite and Patriarch Photius.
- Smaller sub-corpora consist of texts on Geography, Mathematics, History, etc. Of particular interest are annotated editions of works by Thucydides, Homer, Aristotle, etc.
- The content of the manuscripts is also characterised by diversity, covering subjects such as Dogmatics, Hermeneutics, Ascetic, Liturgics, Philosophy and Church Music.
- The manuscripts in the collection were created from the 13th to the 19th century; unfortunately, the dates of a large part of the manuscripts remains unknown.
- The books were printed from the 16th to the 19th century in Venice, Verona, Rome, Genova, Athens, Smyrna, Constantinople, Basel, Vienna, Leipzig, Berlin, Paris, London, St Petersburg, Moscow, Bucharest, Iasi, etc.

The corpus, in its initial state, has drawn the interest of experts in the humanities, who are now able to study the texts from images which are as close as possible to the originals.

3.2 A first step toward availability and awareness

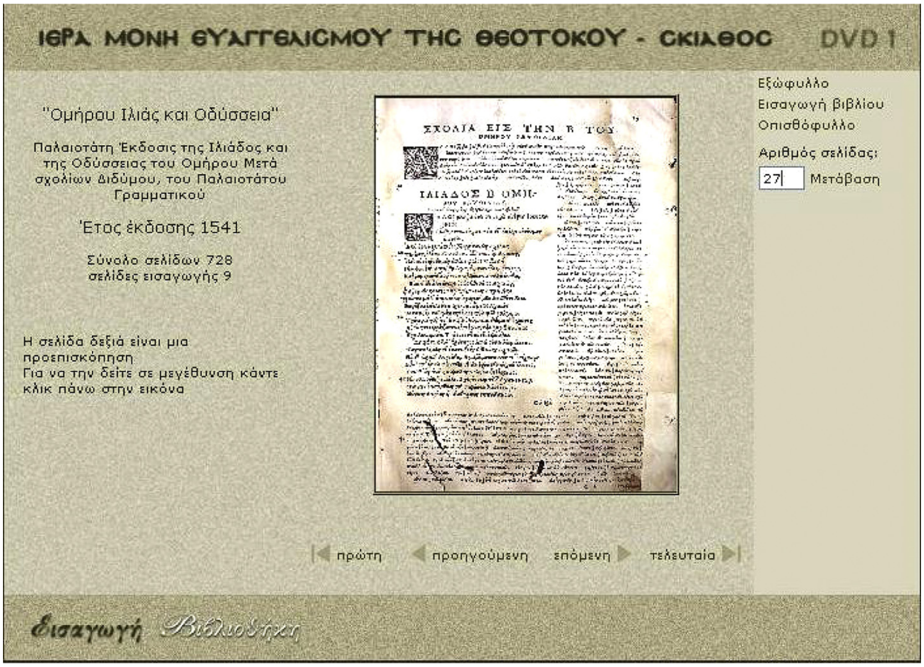
With the digitisation of the collection items and their linking to the database, the first challenge of this undertaking was met; access to the texts became possible without using the originals, thus avoiding further deterioration of their state. However, the availability remained an open issue because the system was less than user-friendly.

To facilitate the availability and awareness of the digitised texts, it was decided to publish a representative subset of printed books on a set of DVDs. User-friendly navigation software to access the digitised material was created, and 59 printed books (corresponding to about 28,000 pages) were processed and published on three DVDs along with the software (Plate 3). The subset of the digitised collection consisting of these 59 books is a representative sample of the items in the library, taking into account their philological, theological and/or historical interest, their rarity, their thematic diversity, etc. Special importance was placed on the user-friendliness of the navigation software because the target group would consist mainly of experts in the humanities who were not particularly familiar with advanced computer use. With the use of the navigation software, the user is able to access a book, inspect each page (Figure 1) and print it for further study. The abovementioned undertaking of digitising the texts and their publication on a set of DVDs was implemented by the Department of Informatics of the Technological Educational Institute of Athens and the company Data and Control Systems Ltd. with the support of the Municipality of Skiathos.


Περιεχόμενα Βιβλία
DVD 1
DVD 2
DVD 3
ΒΙΒΛΙΟΘΗΚΗ ΙΕΡΑΣ ΜΟΝΗΣ ΕΥΑΓΓΕΛΙΣΜΟΥ ΤΗΣ ΘΕΟΤΟΚΟΥ - ΣΚΙΑΘΟΣ
Βιβλιοθήκη
ΙΕΡΑΣ ΜΟΝΗΣ ΕΥΑΓΓΕΛΙΣΜΟΥ ΤΗΣ ΘΕΟΤΟΚΟΥ

Source: Library of the Holy Monastery of the Annunciation of the Virgin, Skiathos, Greece. Photo taken by author

Plate 3.
The 59 books available on DVDs



"Ομήρου Ιλιάς και Οδύσσεια"

Παλαιοτάτη Έκδοσις της Ιλιάδος και της Οδύσσειας του Ομήρου Μετά σχολίων Διδύμου, του Παλαιοτάτου Γραμματικού

Έτος έκδοσης 1541

Σύνολο σελίδων 728
σελίδες εισαγωγής 9

Η σελίδα δεξιά είναι μια προεπισκόπηση
Για να την δείτε σε μεγέθυνση κάντε κλικ πάνω στην εικόνα

Αριθμός σελίδας: 27 Μετάβαση

Figure 1.
Inspection of an annotated edition of Homer's Iliad and Odyssey (Basel, 1541)

Source: Library of the Holy Monastery of the Annunciation of the Virgin, Skiathos, Greece

4. Processing and exploitation

To achieve a quick and efficient content exploitation of the digitised texts, the corpus is subject to processing following a novel approach, which is described below:

4.1 Processing and exploitation

For the effective processing and exploitation of a digitised corpora, an optical character recognition (OCR) step is usually a prerequisite (Bokser, 1992). However, in the case of historical texts, the performance of OCR is strongly affected by factors such as poor paper quality, paper positioning variations, low print contrast and typesetting imperfections. More recently, attempts to improve OCR accuracy for critical editions of classics have been proposed by Boschetti *et al.* (2009). Alternatively, to avoid the use of an OCR system, a method for direct search of keywords into page images (word spotting) in historical texts was proposed by Manmatha and Croft (1997). As this process, which is based on the comparison of entire words rather than individual characters, can become very tedious for large document collections, alternatives and improvements have been proposed such as in studies by Rath and Manmatha (2003), Gatos *et al.* (2006) and Konidaris *et al.* (2007).

4.2 An alternative method for accessing the content of the texts

To process the items of the collection, it was decided to avoid the use of an OCR system because the state of the items would render the results of the OCR procedure highly unreliable. To this end, an alternative method for accessing the content of historical machine-printed texts based on word spotting aided by Natural Language Processing techniques is proposed in Kesidis *et al.* (2011). An initial experimentation is also reported in Kesidis *et al.* (2009).

The method consists of the following five steps:

(1) image preprocessing (Plate 4);

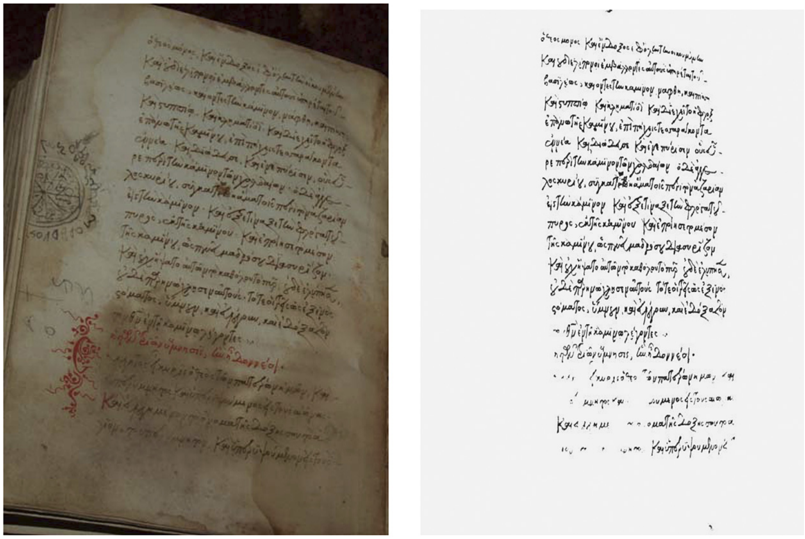
(2) creation of synthetic image words from keywords typed by the user;

(3) word segmentation using dynamic parameters;

(4) feature extraction for each word image; and

(5) a retrieval procedure supported by users' feedback and Natural Language Processing tools.

In this framework, synthetic word images are created from keywords and these images are compared to all the words in the digitised texts. To optimise the accuracy of the results of the query, a user-feedback technique is used. In addition, the aforementioned methods are aided by a set of Natural Language Processing tools to improve accessing and searching the texts. In short, the novelty of the approach lies in the fact that the proposed word image retrieval system combines natural language processing techniques with document image analysis and searching for the improvement in retrieval efficiency. The proposed framework can be part of an information system for the processing, management and accessing the content of the books and manuscripts in the digitised collection.



(a)

(b)

Source: Library of the Holy Monastery of the Annunciation of the Virgin, Skiathos, Greece. Photo taken by author

Plate 4.
(a) Initial state of a manuscript page and (b) manuscript page after image preprocessing

4.3 Improving the efficiency of accessing and searching

An effective search in texts is a Natural Language Processing problem which must take into account the language in which the texts are written. When it comes to historical texts, the situation is more complicated because it is questionable whether tools developed for modern languages are directly applicable to historical texts. In fact, significant differences between the contemporary form of a language and older varieties appearing in the text are observed. Solutions have been proposed for the processing of German historical texts by Ernst-Gerlach and Fuhr (2006), presenting an algorithm for generating search term variants in ancient orthography of German, and by Gotcharek *et al.* (2009), proposing the combination of matching procedures and historical lexica for enabling information retrieval on historical texts.

To improve the efficiency of accessing and searching the corpus, we have developed a set of Natural Language Processing tools which take into account earlier stages of the Greek language. This set of Natural Language Processing tools for older varieties of Greek used for query expansion is described in the following section and comprises:

- a morphological generator which generates all inflected forms of a given keyword, thus providing the user with the ability to use only the base form and to locate all the corresponding inflected word forms in the texts; and
- a synonym dictionary which facilitates access to the semantic content of the texts.

5. Natural language processing tools for older varieties of Greek

Until recently, attempts at a computational processing of older varieties of the Greek language have focused on Ancient Greek. Tools, such as morphological analysers, have been proposed either as pedagogical tools (Packard, 1973), or for research purposes, such as the rule-based Morpheus morphological analyser (Crane, 1991), which is extensively used in the Perseus digital library (Crane, 1996). More recently, a data-driven morphological analyser for Ancient Greek in a machine-learning framework was proposed (Lee, 2008). However, such tools would not be sufficient for a processing of the language in a transitional period such as the one covered by the texts in the collection.

The tools that are described in subsections 5.2 and 5.3 were developed to deal with a sub-part of the corpus, containing books that were printed during the 17th and 18th centuries. That particular period was chosen for this first implementation because the language in which the texts are written reflects an early stage of Modern Greek and is indicative of the evolution of the Greek language, as it incorporates elements from Ancient, Medieval and Modern Greek. An initial version of these tools is described in Kesidis *et al.* (2009).

5.1 The morphological generator

The implementation of the morphological generator is based on finite state technology which has been extensively used for the morphological processing of a large number of languages (Karttunen and Oflazer, 2000). In particular, morphological phenomena in Modern Greek, such as inflection (Sgarbas and Kokkinakis, 1995) and derivation and compounding (Ralli and Galiotou, 2004), have been extensively processed within the finite state framework. However, the aforementioned approaches are not directly applicable to the implementation of inflectional morphology of earlier stages of the language mainly for two reasons:

(1) the set of inflectional affixes has diminished and/or have been subject to change in the course of the evolution of the language; and

(2) certain grammatical cases appearing in historical texts, such as the dative case, have disappeared in the current stage of Greek.

Therefore, a morphological generator for early Modern Greek embedding the Stuttgart finite state transducer (SFST) tools (Schmid, 2005) was implemented. The SFST collection of software tools for the generation, manipulation and processing of finite-state transducers specified by means of the SFST programming language was developed by the Institute for Natural Language Processing, University of Stuttgart[4]. The implementation is based on an extensive linguistic analysis of the nominal inflection phenomenon in this transitional stage of the language (Kesidis *et al.*, 2009). The tool provides the user with the following three possibilities:

(1) insert/edit/delete a keyword;

(2) assign its inflectional class through the selection of the appropriate class representative; and

(3) perform a generation of all inflected word forms according to the rules of the appropriate inflectional class (Figure 2).

Class representatives enable the use of the system without specific knowledge of the implemented morphological rules. Knowledge of the language at the native speaker level is sufficient. Therefore, the user is able to enrich the set of keywords simply by defining their correspondence to the representatives of inflectional classes and automatically generate all inflected word forms without the explicit use of the morphological rules.

5.2 The synonym dictionary

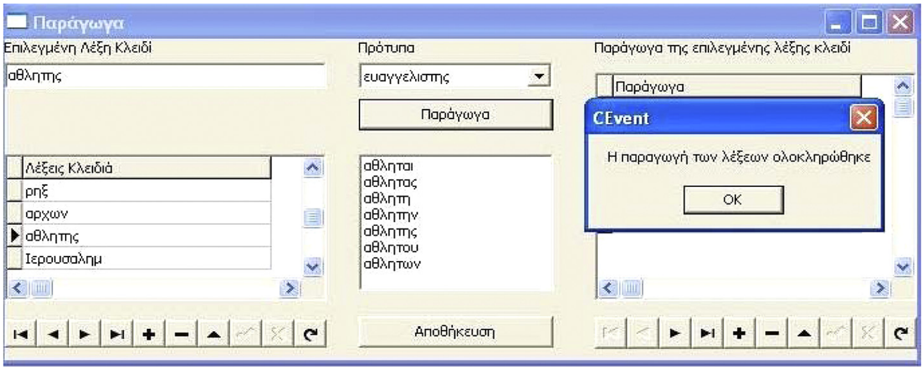
The access to the semantic content of texts is, in general, achieved with the use of semantic networks such as WordNet (Voorhees, 1998). However, the usage of the semantic network is often questioned with regard to improvement in the performance of the query expansion procedure because it often entails an unnecessary amount of ambiguity. Therefore, to further enrich the searching procedure and enable the access to the semantic content of the texts, we have developed a synonym dictionary. The synonymy embedded in the dictionary is based on a relevance feedback technique, as its functionality is equivalent to that of a semantic network. We have followed the methodology proposed by Turcato *et al.* (2000), in particular in ranking the synonymy relations in terms of their weight in the religious–historical domain to associate them with a score that estimates how often a relation is used in the specific domain. Relations that are irrelevant to the domain are not inserted in the synonym database. In our approach, the synonyms of a word are ranked according to their relevance to the original word on a scale from 1 to 10, the value of 1 denoting the highest relevance to the original word.

The tool provides the user with the following two possibilities:

(1) add, edit and delete a word in the dictionary, accompanied by up to five synonyms. Each synonym is given a weighted value according to the relevance of the synonym to the original word. The values range from 1 (the highest relevance to the word) to 10; and

(2) look up a word and its synonyms (Figure 3).

The system provides the user with the ability to extract all inflected forms and weighted synonyms of a given base form of a word. The ensemble of extracted word forms is used for expanding the query during the word spotting process (Figure 4).



Παράγωγα

Επιλεγμένη Λέξη Κλειδί: αθλητης

Λέξεις Κλειδιά
ρηξ
αρχων
αθλητης
Ιερουσαλημ

Πρότυπα: ευαγγελιστης

αθληται
αθλητας
αθλητη
αθλητην
αθλητης
αθλητου
αθλητων

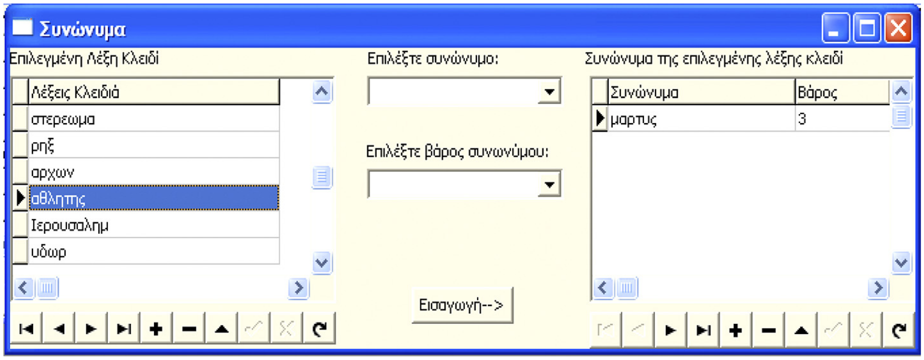
Παράγωγα της επιλεγμένης λέξης κλειδί

Παράγωγα

CEvent

Η παραγωγή των λέξεων ολοκληρώθηκε

Figure 2.
Generation of inflected forms for the keyword "*αθλητης*"("*athlete*") according to the pattern "*ευαγγελιστης*" ("*evangelist*")



Συνώνυμα

Επιλεγμένη Λέξη Κλειδί

Λέξεις Κλειδιά
στερεωμα
ρηξ
αρχων
αθλητης
Ιερουσαλημ
υδωρ

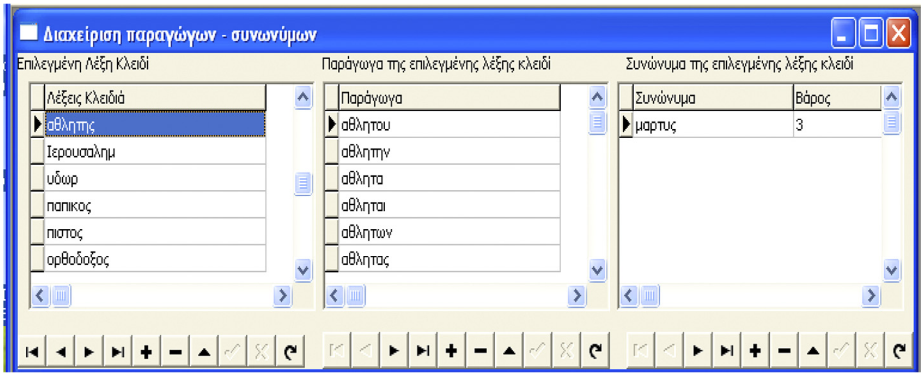
Επιλέξτε συνώνυμο:

Επιλέξτε βάρος συνωνύμου:

Συνώνυμα της επιλεγμένης λέξης κλειδί

Συνώνυμα	Βάρος
μαρτυς	3

Figure 3.
Synonym dictionary lookup for keyword "*αθλητης*" ("*athlete*"). In the religious domain, the word "*μαρτυς*" ("*martyr/witness*") is its synonym with a weighted value of 3



Διαχείριση παραγώγων - συνωνύμων

Επιλεγμένη Λέξη Κλειδί

Λέξεις Κλειδιά
αθλητης
Ιερουσαλημ
υδωρ
παπικος
πιστος
ορθοδοξος

Παράγωγα της επιλεγμένης λέξης κλειδί

Παράγωγα
αθλητου
αθλητην
αθλητα
αθληται
αθλητων
αθλητας

Συνώνυμα της επιλεγμένης λέξης κλειδί

Συνώνυμα	Βάρος
μαρτυς	3

Figure 4.
Extraction of inflected word forms and weighted synonyms for the keyword "*αθλητης*" ("*athlete*"). The ensemble of the extracted words will be used for query expansion

5.3 Experimentation

An experiment with an initial version of the tools described above is reported in Kesidis *et al.* (2011). It was performed on 110 digitised pages of historical/religious texts printed during the 17th and 18th centuries. The query set consisted of 32 keywords and was expanded with inflected word forms produced by the morphological generator and the synonym dictionary. The results of the experiment have shown that the use of the automatically generated word forms in the word spotting procedure has led to significant improvement in the retrieval performance when compared to queries where only the base form of a keyword is used.

6. Conclusions

In this paper, an attempt to build and process a Greek digital historical corpus was described. The corpus is the outcome of a digitisation project, aiming at contributing to the preservation and availability of manuscripts created from the 13th to the 19th centuries and books printed from the 16th to the 19th centuries. The first phase of this ongoing project comprises:

- the creation of the infrastructure for the inventory and digitisation of the collection items;
- the digitization of the texts; and
- the publication of a representative number of digitised texts on a set of DVDs.

The exploitation and processing of the texts are performed without the use of an OCR system. In particular, access to the content of the digitised texts is performed by searching directly in the digitised texts with the help of Natural Language Processing techniques, which leads to a significant improvement in the retrieval performance. In its present state, the use of the corpus facilitates the work of theologians, historians, philologists, paleographers, etc. and, at the same time, prevents the original collection items from further damage. In addition, the Natural Language Processing tools presented in this paper have given insight as to the characteristics of the Greek language during this transitional period and constitute a first step towards the development of computational tools for the study of the diachronic evolution of the language.

Future work involves:

- the enhancement and improvement of the aforementioned methods to process the complete collection of digitised texts;
- the adoption of best practice strategies in digitising to generate metadata and to maintain and preserve the digital corpus;
- the preservation of the original collection of books and manuscripts; and
- further development of Natural Language Processing tools to take into account the diachronic evolution of the Greek language.

Notes

1. www.perseus.tufts.edu
2. dlab.phs.uoa.gr/ellinomnimon.htm
3. http://anemi.lib.uoc.gr
4. www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html

References

Anthony (Former Bishop of Elia) (1961), "The catalog of manuscripts and printed books of the holy monastery of the annunciation of the virgin in skiathos (Part I)", *Theologia*, Vol. 31 No. 4.

Anthony (Former Bishop of Elia) (1962a), "The catalog of manuscripts and printed books of the holy monastery of the annunciation of the virgin in skiathos (Part II)", *Theologia*, Vol. 32 No. 2.

Anthony (Former Bishop of Elia) (1962b), "The catalog of manuscripts and printed books of the holy monastery of the annunciation of the virgin in skiathos (Part III)", *Theologia*, Vol. 32 No. 4.

Balk, H. and Conteh, A. (2011), "IMPACT: centre of competence in text digitisation", in *Workshop on Historical Document Imaging and Processing (HIP '11), Beijing*, ACM Press, New York, NY, pp. 155-160.

Bokser, C. (1992), "Omnidocument technologies", in *Proceedings of the IEEE*, Vol. 80 No. 7, pp. 1066-1078.

Boschetti, F., Romanello, M., Babeu, A., Bamman, D. and Crane, G. (2009), "Improving OCR accuracy for classical critical editions", in Agosti, M., Borbinha, J.L., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (Eds), *Proceedings of ECDL 2009*, LNCS 5714, Springer, Berlin, pp. 156-167.

Crane, G. (1991), "Generating and parsing classical Greek", *Literary and Linguistic Computing*, Vol. 6 No. 4, pp. 243-245.

Crane, G. (1996), *Perseus 2.0: Interactive Sources and Studies on Ancient Greece*, Yale University Press, New Haven, CT.

Ernst-Gerlach, A. and Fuhr, N. (2006), "Generating term variants for text collections with historic spellings", in *Proceedings of the 28th European Conference on Information Retrieval Research (ECIR 2006)*, Springer, Berlin.

Frey, J. and Lopes, A. (2011), "Digitization standards", paper presented at the *LIBER-EBLIDA Digitisation Workshop 2011*, *The Hague*.

Gatos, B., Pratikakis, I. and Perantonis, S.J. (2006), "Adaptive degraded document image Binarization", *Pattern Recognition*, Vol. 39 No. 3, pp. 317-327.

Gotcharek, A., Neumann, A., Reffle, U., Ringlstetter, C. and Schulz, K. (2009), "Enabling information retrieval on historical document collections – the role of matching procedures and special lexica", in Lopresti, D., Roy, S., Schulz, K., Subramaniam, L.V. (eds), *Proceedings of the 3rd Workshop on Analytics for Noisy Unstructured Data (AND'09), Barcelona*, ACM Press, New York, NY, pp. 69-76.

Karttunen, L. and Oflazer, K. (2000) (2000), "Special issue on finite state methods in NLP", *Computational Linguistics*, Vol. 26 No. 1.

Kesidis, A.L., Galiotou, E., Gatos, B., Pratikakis, I., Manolessou, I. and Ralli, A. (2009), "Accessing the content of Greek historical documents", in Lopresti, D., Roy, S., Schulz, K., Subramaniam, L.V., (Eds), *Proceedings of the 3rd Workshop on Analytics for Noisy Unstructured Data (AND'09)*, *Barcelona*, ACM Press, New York, NY, pp. 55-62.

Kesidis, A.L., Galiotou, E., Gatos, B. and Pratikakis, I. (2011), "A word-spotting framework for historical machine-printed documents", *International Journal on Document Analysis and Recognition (IJDAR)*, Vol. 14 No. 2, pp. 131-144.

Konidaris, T., Gatos, B., Ntzios, K., Pratikakis, I. and Theodoridis, S. (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 Historical Documents*, Vol. 9 Nos 2/4, pp. 167-177.

Lee, J. (2008), "A nearest-neighbor approach to the automatic analysis of Ancient Greek morphology", in *Proceedings of the 12th Conference on Computational Natural Language Learning, Manchester*, pp. 127-134.

Manmatha, R. and Croft, W.B. (1997), "A draft of word spotting: Indexing handwritten manuscripts", in Maybury, M.T. (Ed.), *Intelligent Multimedia Information Retrieval*, MIT Press, Cambridge, MA, pp. 43-64.

Packard, D.W. (1973), "Computer-assisted morphological analysis of Ancient Greek", in *Proceedings of the 5th Conference on Computational Linguistics, Pisa*.

Ralli, A. and Galiotou, E. (2004), "Greek compounds: a challenging case for the parsing techniques of PC-KIMMO v.2", *International Journal on Computational Intelligence*, Vol. 1 No. 2, pp. 152-162.

Rath, T.M. and Manmatha, R. (2003), "Features for word spotting in historical documents", in *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03)*, Edinburgh, Scotland, pp. 218-222.

Schmid, H. (2005), "A programming language for finite-state transducers", in *Proceedings of the 5th International Workshop on Finite State Methods in Natural Language Processing (FSMNLP 2005), Helsinki*.

Sgarbas, K. and Kokkinakis, N.G. (1995), "A PC-KIMMO based morphological description of Modern Greek", *Literary and Linguistic Computing*, Vol. 10 No. 2, pp. 189-201.

Sotiropoulos, D., Galiotou, E. and Skourlas, C. (2007), "Application of a word-alignment algorithm to bilingual Greek-Latin documents", in *Proceedings of the 7th WSEAS International Conference on Applied Computer Science (ACS'07), Venice*, pp. 238-241.

Sotošek, K.S. (2011), *Best Practice in Library Digitization*, Europeana Travel, Deliverable D2.2, Ljubljana.

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

Varveris, A. and Giannakopoulos, G. (2009), "Managing libraries from a distance: the paradigm of "Neophytos Doukas" municipal library", in *Proceedings of the International Conference on Tourism Development and Management*, pp. 614-619.

Voorhees, E.M. (1998), "Using WordNet for text retrieval", in Fellbaum, C. (Ed.), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, pp. 285-303.

Corresponding author
Eleni Galiotou can be contacted at: egali@teiath.gr