GRPOLY-DB: An Old Greek Polytonic Document Image Database

Basilis Gatos[1], Nikolaos Stamatopoulos[1], Georgios Louloudis[1], Giorgos Sfikas[1], George Retsinas[1], Vassilis Papavassiliou[2], Fotini Simistira[2] and Vassilis Katsouros[2]

[1]Computational Intelligence Laboratory, Institute of Informatics and Telecommunications
National Center for Scientific Research "Demokritos"
GR-153 10 Agia Paraskevi, Athens, Greece
{bgat, nstam, louloud, sfikas, georgeretsi}@iit.demokritos.gr

[2]Institute for Language and Speech Processing (ILSP)
Athena Research and Innovation Center
GR-151 25 Maroussi, Athens, Greece
{vpapa, fotini, vsk}@ilsp.athena-innovation.gr

***Abstract*—Recognition of old Greek document images containing polytonic (multi accent) characters is a challenging task due to the large number of existing character classes (more than 270) which cannot be handled sufficiently by current OCR technologies. Taking into account that the Greek polytonic system was used from the late antiquity until recently, a large amount of scanned Greek documents still remains without full text search capabilities. In order to assist the progress of relevant research, this paper introduces the first publicly available old Greek polytonic database GRPOLY-DB for the evaluation of several document image processing tasks. It contains both machine-printed and handwritten documents as well as annotation with ground-truth information that can be used for training and evaluation of the most common document image processing tasks, i.e., text line and word segmentation, text recognition, isolated character recognition and word spotting. Results using several representative baseline technologies are also presented in order to help researchers evaluate their methods and advance the frontiers of old Greek document image recognition and word spotting.**

Keywords—performance evaluation; benchmarking, Old Greek poytonic characters; character recognition; word spotting

I. INTRODUCTION

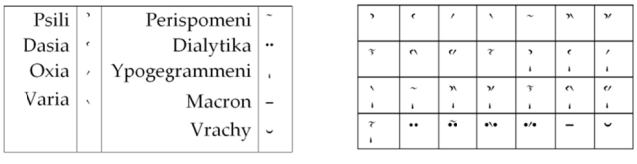
Several databases have emerged during the last decades (e.g. CENPARMI [1], CEDAR [2], IAM [3], George Washington [4]) in order to help researchers compare and evaluate the performance of several document image processing tasks including handwriting segmentation, text recognition, graphics recognition and word spotting. Public databases help researchers to advance the state-of-the-art since they permit a fair and objective comparison under a common scenario. An overview of existing datasets and annotations for document analysis and recognition can be found at [5].

In this paper, we introduce the first publicly available old Greek polytonic database GRPOLY-DB. The existence of a large number of character classes (more than 270) makes recognition of old Greek document images a challenging task that cannot be handled sufficiently by current OCR technologies. A large amount of scanned Greek documents still remains without full text search capabilities taking into account that the Greek polytonic system was used from the late antiquity until recently (1982). GRPOLY-DB contains both machine-printed and handwritten documents as well as annotation with ground-truth information that can be used for training and evaluation of the four most common document image processing tasks, i.e., text line and word segmentation, text recognition, isolated character recognition and word spotting. In order to help researchers evaluate their methods and advance the frontiers of old Greek document image recognition and word spotting, we also provide results using several representative baseline technologies.

The remainder of the paper is organized as follows. In Section II, the properties of the old Greek polytonic documents are discussed. In Section III, an overview of the GRPOLY-DB is presented while the workflow used to create GRPOLY-DB is presented in Section IV. Evaluation results using several baselines methods are presented in Section V. Finally, conclusions are drawn in Section VI.

II. PROPERTIES OF THE OLD GREEK POLYTONIC DOCUMENTS

The Greek polytonic system includes 9 diacritic marks (Fig.1a). Some of these marks are combined and as a result we have a total of 28 different diacritic mark combinations that may appear above or below Greek characters (Fig. 1b). The total number of Greek characters is 49 (25 lower case and 24 upper case). These characters are combined with the diacritic marks and as a result we have more than 270 character classes (see Fig. 2).



(a) (b)

Fig. 1. Old Greek polytonic system: a) Diacritic marks and b) their combinations.

ἀ ἁ ἂ ἃ ἄ ἅ ἆ ἇ ὰ ᾀ ᾁ ᾂ ᾃ ᾄ ᾅ ᾆ ᾇ ᾲ ᾳ ᾴ ᾶ ᾷ ᾰ ᾱ
Ἀ Ἁ Ἂ Ἃ Ἄ Ἅ Ἆ Ἇ Ὰ Ᾰ Ᾱ

Fig. 2. Greek polytonic characters based on “α” and “A”.

III. Overview of the GRPOLY-DB

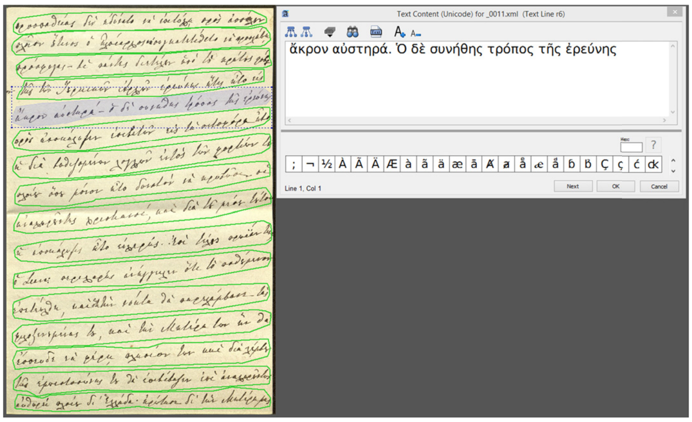
GRPOLY-DB can be downloaded from [6] and consists of four subsets that have been semi-automatically annotated with ground-truth information at different levels using the PAGE (Page Analysis and Ground-Truth Elements) format [7]. An overview of GRPOLY-DB subsets is presented in Table I. For every segmentation level, the correspondence with the polytonic text is provided (see Fig.3). In this section, information about all GRPOLY-DB subsets is provided.

TABLE I. Overview of GRPOLY-DB subsets

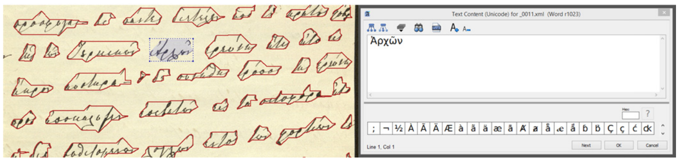
	Date	Pages	Number of Text Lines with GT	Number of Words with GT	Number of Characters with GT
GRPOLY-DB-Handwritten	1838-1916	46	693	4939	-
GRPOLY-DB-MachinePrinted-A	1950-1965	5	691	4998	28591
GRPOLY-DB-MachinePrinted-B (1-4)	1864	6	653	5895	30533
	1931	5	522	4473	22923
	1953	18	1673	13076	72750
	1977	4	374	3340	16714
GRPOLY-DB-MachinePrinted-C	1912	315	10478	65875	-
Total		**399**	**15084**	**102596**	**171511**

A. GRPOLY-DB-Handwritten

This part contains 46 color page images from a historical manuscript written by Sofia Trikoupi (1838-1916) [8]. The corresponding ground-truth contains segmentation at text line and word level (see Fig. 3).



(a)

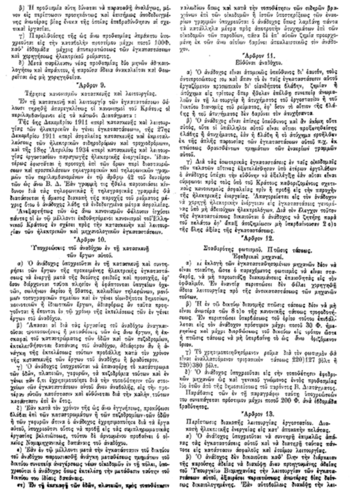


(b)

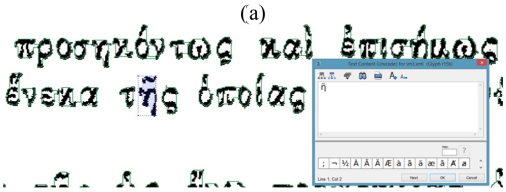
Fig. 3. Examples of GRPOLY-DB-Handwritten ground-truth regions at text line (a) and word (b) level.

B. GRPOLY-DB-MachinePrinted-A

It consists of 5 binary page images obtained from the Hellenic National Printing House which is responsible for publishing digital copies of the Laws and Presidential decrees of the Greek State [9]. The publication year of these documents ranges between 1950 and 1965. A document image sample is shown in Fig. 4. The corresponding ground-truth contains segmentation at text line, word and character level.



(a)



(b)

Fig. 4. Example of a GRPOLY-DB-MachinePrinted-A image (a) and ground-truth at character level (b).

C. GRPOLY-DB-MachinePrinted-B

This subset consists of 33 grayscale page images from the parliament session proceedings dated from 1864 to 1977 originating from the archive of the Hellenic Parliament [10] (see Fig. 5). The corresponding ground-truth contains segmentation at text line, word and character level. It is further divided based on the corresponding period to B1 (6 pages dated at 1864), B2 (5 pages dated at 1931), B3 (18 pages dated at 1953) and B4 (4 pages dated at 1977). These pages correspond to speeches of four Greek politicians (Saripolos in 1864 and Venizelos in 1931, Markezinis in 1953 and Vlahou in 1977).

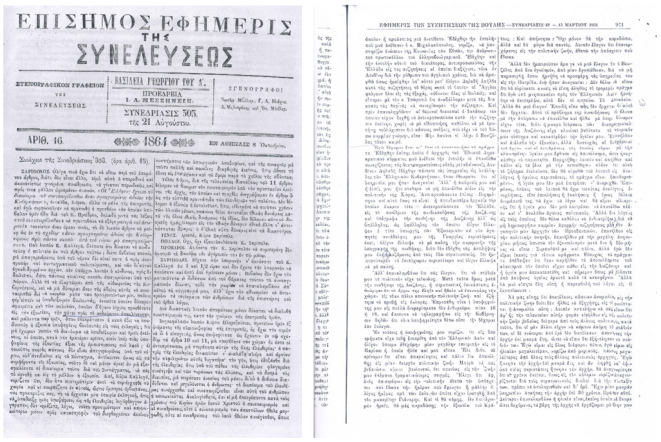


Fig. 5. Examples of GRPOLY-DB-MachinePrinted-B images.

D. GRPOLY-DB-MachinePrinted-C

It contains 315 color page images from the Appian's Roman History Books I – VIII [11] and the corresponding ground-truth segmentation at text line and word level. It is a set of better quality compared to other machine printed GRPOLY-DB sets (see Fig. 6).

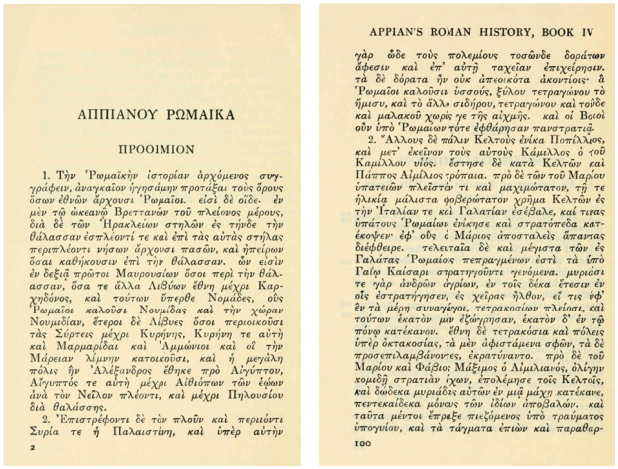


Fig. 6. Examples of GRPOLY-DB-MachinePrinted-C images.

IV. Creation of GRPOLY-DB

For the creation of the Polyton-DB, we used the original images of all pages as well as the corresponding transcription. We first binarized [12] the original images if required and then applied layout analysis and segmentation processes [13]. At a next step, several users were involved in order to correct the segmentation results using the Aletheia framework [14]. An automatic transcript mapping procedure was applied in order to assign the text information to the corresponding text line [15]. This procedure was also verified and corrected by a set of users.

V. Evaluation on GRPOLY-DB

In this section, we evaluate representative baseline document image processing techniques on GRPOLY-DB. We also demonstrate how this database can be used for evaluating the most common document imaging tasks, i.e., text line and word segmentation, text recognition, isolated character recognition and word spotting. For all cases that the database is split to training and test parts, the corresponding partitioning can be found at [6].

A. Text line segmentation

All subsets of GRPOLY-DB were used to evaluate the performance of the following state-of-the-art text line segmentation algorithms: Based on Shredding [16] and on Hough transform [17]. We followed the evaluation protocol used in handwriting segmentation contests [18] in terms of Detection Rate (*DR*), Recognition Accuracy (*RA*) and F-Measure (*FM*). Comparative results are shown in Table II.

TABLE II. Text Line Segmentation Results

		DR (%)	*RA (%)*	*FM(%)*
Shredding based method [16]	GRPOLY-DB-Handwritten	92.35	80.60	86.08
	GRPOLY-DB-MachinePrinted-A	89.2	93.91	91.54
	GRPOLY-DB-MachinePrinted-B	95.44	95.82	95.63
	GRPOLY-DB-MachinePrinted-C	93.48	96.69	95.06
	TOTAL	**93.66**	**95.52**	**94.58**
Hough transform based [17]	GRPOLY-DB-Handwritten	96.68	94.23	95.44
	GRPOLY-DB-MachinePrinted-A	95.95	97.64	96.79
	GRPOLY-DB-MachinePrinted-B	98.23	96.91	97.56
	GRPOLY-DB-MachinePrinted-C	87.63	93.86	90.64
	TOTAL	**90.69**	**94.74**	**92.67**

B. Word segmentation

Using the previously described evaluation protocol, we evaluated the following state-of-the-art word segmentation algorithms: Based on sequential clustering [19] and on Gaussian mixtures [20]. The corresponding results are shown in Table III.

TABLE III. Word Segmentation Results

		DR (%)	*RA (%)*	*FM(%)*
Sequential clustering [19] based method	GRPOLY-DB-Handwritten	76.41	84.56	80.28
	GRPOLY-DB-MachinePrinted-A	94.60	92.96	93.77
	GRPOLY-DB-MachinePrinted-B	93.62	93.59	93.60
	GRPOLY-DB-MachinePrinted-C	96.15	96.83	96.49
	TOTAL	**94.46**	**95.24**	**94.85**
Gaussian mixtures [20] based method	GRPOLY-DB-Handwritten	82.83	85.86	84.32
	GRPOLY-DB-MachinePrinted-A	92.72	88.23	90.42
	GRPOLY-DB-MachinePrinted-B	92.05	86.52	89.20
	GRPOLY-DB-MachinePrinted-C	97.35	93.76	95.52
	TOTAL	**95.04**	**91.21**	**93.08**

C. Isolated character recognition

A 5-fold cross-validation was applied in order to evaluate the recognition of isolated characters of subset GRPOLY-DB-MachinePrinted-B. We first selected all characters belonging to classes with at least 30 instances (125 classes). Two different scenarios were defined. According to the first scenario (SC-1), all instances were used (143051 instances) while at the second scenario (SC-2), only 30 randomly selected instances per class were used (3750 instances). We evaluated two state-of-the-art character recognition techniques based on HoG features [21] combined with an SVM classifier and adaptive windows features [22] combined with a KNN classifier. The corresponding results are shown in Table IV.

TABLE IV. ISOLATED CHARACTER RECOGNITION RESULTS

		Recognition Accuracy (%)
HoG features [21] – SVM	SC-1	98.37
	SC-2	92.00
Adaptive Windows features [22] - KNN	SC-1	97.71
	SC-2	88.69

D. Text recognition

We evaluated the OCR recognition performance at character and word levels using (a) the open source OCR engine of Tesseract [23] and (b) the commercial OCR FineReader Engine v.11 [24] on the dataset GRPOLY-DB-MachinePrinted-B. Several text blocks that do not contain non-Greek symbols and correspond to 2835 text lines were cropped and used as input in order to test both recognition engines. For Tesseract no training was necessary, as we used the model for Greek polytonic built by Nick White [25]. For the ABBYY FineReader Engine we used 367 text lines of GRPOLY-DB-MachinePrinted-B that do not belong to the test set in a way so that each target character-class to appear at least 5 times. We semi-automatically segmented the selected text line images into character images and used the training utility of the ABBYY FineReader engine SDK to create the respective characters' models (we created 4 recognition databases that correspond to GRPOLY-DB-MachinePrinted-B 1-4 in order to be used with the corresponding testing sets). In addition, we have built a dictionary for Katharevousa (a form of the Greek language in the early 19th century) with the use of texts from the Thesaurus Linguae Graecae corpus [26]. The evaluation results recorded concerning error rates on character and word level are presented in Table V.

E. Word spotting

We ran word spotting trials using two well-known learning-free, segmentation-based methods, adaptive windows [22] and profiles [4]. Both methods are suitable for Query-by-example (QBE) word spotting. Adaptive windows create a fixed-length descriptor for each segmented word image that can in turn be compared with descriptors of other images in the database using the Euclidean distance. The size of profile features is dependent on the length of the input word image, and comparison between feature vectors can be achieved using Dynamic Time Warping (DTW) which optimizes correspondence between matching feature components using dynamic programming. The input for our experiments is all cropped, binarized word images from GRPOLY-DB. Evaluation is performed using a fixed set of queries for each dataset. The query descriptor is matched against all other descriptors in the dataset, and image distances that fall under a variable threshold are considered matches. For the purpose of evaluation, we are not interested on a specific distance threshold, but we calculate average Precision [27] as a metric on all possible thresholds. Mean average Precision (MAP) is then calculated as the mean over all query evaluation results. In our numerical experiments, we use the implementation of the Text REtrieval Conference (TREC) community by the National Institute of Standards and Technology (NIST) [28]. In order to choose queries with a well-defined criterion, we follow the rationale of the recent word spotting contest [27] and define our query list on the basis of transcription length and number of instances in the whole dataset. For each of our subset, we add to our query list all words that have more than 6 characters and appearing more than 5 times. Specifically for GRPOLY-DB-Handwritten we apply a less strict criterion due to the small size of this set, and use words with more than 5 characters and 4 appearances. The evaluation results concerning MAP for each of our datasets are presented in table VI.

TABLE V. CHARACTER ERROR RATES (CER) AND WORD ERROR RATES (WER) FOR TEXT RECOGNITION

	Tesseract		***ABBYY FineReader***	
	CER (%)	***WER (%)***	***CER (%)***	***WER (%)***
GRPOLY-DB-MachinePrinted-B1	28.41	71.71	23.6	46.69
GRPOLY-DB-MachinePrinted-B2	22.29	66.71	15.28	55.54
GRPOLY-DB-MachinePrinted-B3	31.13	71.36	19.70	48.61
GRPOLY-DB-MachinePrinted-B4	42.30	77.61	14.34	42.51
Average	**30.37**	**71.43**	**19.20**	**48.60**

TABLE VI. MEAN AVERAGE PRECISION (MAP) RESULTS FOR WORD SPOTTING

	Number of query words	***MAP(%)***	
		Adaptive windows [22]	*Profiles+ DTW [4]*
GRPOLY-DB-Handwritten	21	40.4	56.2
GRPOLY-DB-MachinePrinted-A	35	68.2	89.8
GRPOLY-DB-MachinePrinted-B	103	57.8	62.0
GRPOLY-DB-MachinePrinted-C	363	79.0	87.7
Average		**61.35**	**73.93**

VI. Conclusions

In this paper, the first publicly available old Greek polytonic database GRPOLY-DB is introduced. It contains both machine-printed and handwritten documents as well as annotation with ground-truth information at several levels (text line, word, character level). For every segmentation level, the correspondence with the polytonic text is also provided. Representative state-of-the-art methods are used for applying the most common document image processing tasks, i.e., text line and word segmentation, text recognition, isolated character recognition and word spotting, on GRPOLY-DB. The following indicative evaluation results have been recorded: Text line segmentation using a shredding based method [16]: **94.58%** (F-Measure), word segmentation using a sequential clustering [19] based method: **94.85%** (F-Measure), isolated character recognition using HoG features [21] and SVM classifier: **98.37%** (Recognition Accuracy), text recognition using FineReader Engine v.11 [24]: **19.20%** (Character Error Rate), word spotting using Profiles and DTW [4]: **73.93%** (Mean average Precision).

Acknowledgment

This work has been supported by the OldDocPro project (ID 4717) funded by the GSRT .

References

[1] C.Y. Suen, C. Nadal, R. Legault, T.A. Mai and L. Lam, “Computer recognition of unconstrained handwritten numerals”, Proc. IEEE, vol. 80, no. 7, pp. 1162-1180, 1992.

[2] J. Hull, “A database for handwritten text recognition research”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 5, pp. 550-554, 1994.

[3] U.V. Marti and H. Bunke, “The IAM-database: an English sentence database for offline handwriting recognition”, International Journal on Document Analysis and Recognition (IJDAR), vol. 5, no. 1, pp. 39-46, 2002.

[4] T.M. Rath and R. Manmatha, “Word spotting for historical documents”, International Journal on Document Analysis and Recognition (IJDAR), vol. 9, no. 2, pp. 139–152, 2007.

[5] E. Valveny, “Datasets and Annotations for Document Analysis and Recognition”, Handbook of Document Image Processing and Recognition, D. Doermann, K. Tombre (eds.), Springer-Verlag London 2014.

[6] http://www.iit.demokritos.gr/~nstam/GRPOLY-DB

[7] S. Pletschacher and A. Antonacopoulos, “The PAGE (Page Analysis and Ground-Truth Elements) Format Framework”, 20th International Conference on Pattern Recognition (ICPR 2010), pp. 257-260, 2010

[8] S. Trikoupi, “Η Μήτηρ μας Αικατερίνη Τρικούπη, το γένος Νικολάου Μαυροκορδάτου (1800-1871)”, Library of the Hellenic Parliament, Athens 2012, ISBN 978-960-560-110-2.

[9] http://www.et.gr/

[10] http://www.hellenicparliament.gr/en/

[11] Appian, Appian's Roman history in four volumes. Vol. 1 / Books I-VIII, Heinemann, 1912

[12] B. Gatos, I. Pratikakis and S. J. Perantonis, "Adaptive Degraded Document Image Binarization", Pattern Recognition, Vol. 39, pp. 317-327, 2006.

[13] B. Gatos, G. Louloudis and N. Stamatopoulos, “Segmentation of Historical Handwritten Documents into Text Zones and Text Lines”, 14th International Conference on Frontiers in Handwriting Recognition (ICFHR'14), pp. 464-469, Creta, Grecce, September 2014.

[14] C. Clausner, S. Pletschacher, A. Antonacopoulos, “Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments”, 11th International Conference on Document Analysis and Recognition (ICDAR2011), pp. 48-52, Beijing, China, September 2011.

[15] N. Stamatopoulos, G. Louloudis and B. Gatos, “Efficient Transcript Mapping to Ease the Creation of Document Image Segmentation Ground Truth with Text-Image Alignment”, 12th International Conference on Frontiers in Handwriting Recognition (ICFHR'10), pp. 226-231, Kolkata, India, November 2010.

[16] A. Nicolaou and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines", 10th International Conference on Document Analysis and Recognition (ICDAR'09), pp. 626-630, Barcelona, Spain, July 2009.

[17] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis, "Text line detection in handwritten documents", Pattern Recognition, Vol. 41, Issue 12, pp. 3758-3772, 2008.

[18] N. Stamatopoulos, G. Louloudis, B. Gatos, U. Pal and A. Alaei, “ICDAR2013 Handwriting Segmentation Contest”, 12th International Conference on Document Analysis and Recognition (ICDAR'13), pp. 1402-1406, Washington DC, USA, August 2013.

[19] S.H. Kim, S. Jeong, G.- S. Lee, C.Y. Suen, "Word segmentation in handwritten Korean text lines based on gap clustering techniques", Proc. 6th Int’l Conf. on Document Analysis and Recognition (ICDAR’01), 2001, pp. 189-193.

[20] G. Louloudis, B. Gatos, I. Pratikakis and C. Halatsis, "Text line and word segmentation of handwritten documents", Pattern Recognition, 42(12), pp. 3169-3183, 2009.

[21] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", IEEE Conference Computer Vision and Pattern Recognition, pp. 886-893, 2005.

[22] B. Gatos, A. Kesidis and A. Papandreou, “Adaptive Zoning Features for Character and Word Recognition”, 11th International Conference on Document Analysis and Recognition (ICDAR'11), pp. 1160 - 1164, Beijing, China, 2011.

[23] https://code.google.com/p/tesseract-ocr/

[24] http://www.abbyy.com.gr/ocr_sdk/

[25] Nick White, “Training Tesseract for Ancient Greek OCR,” Εὔτυπον, No 28-29 - October 2012 (retrieved from http://eutypon.gr/eutypon/pdf/e2012-29/e29-a01.pdf).

[26] http://www.tlg.uci.edu/

[27] I. Pratikakis, K. Zagoris, B. Gatos, G. Louloudis, N. Stamatopoulos, ICFHR 2014 Competition on Handwritten KeyWord Spotting (H-KWS 2014), 14th International Conference on Frontiers in Handwriting Recognition (ICFHR’14), pp. 814 – 819, Chersonissos, Greece, 2014.

[28] TREC NIST (2013)

Available: http://trec.nist. gov/pubs/trec16/appendices/measures.pdf