Преглед НЦД 21 (2012), 117–126

Ines Jerele
Library Research Centre,
National and University Library, Slovenia
Tomaž Erjavec
Department of Knowledge Technologies,
Jožef Stefan Institute, Slovenia
Daša Pokorn
Digital Library Development and Strategy Division,
National and University Library, Slovenia
Alenka Kavčič-Čolić
Research and Development Division,
National and University Library, Slovenia

OPTICAL CHARACTER RECOGNITION OF HISTORICAL TEXTS: END-USER FOCUSED RESEARCH FOR SLOVENIAN BOOKS AND NEWSPAPERS FROM THE 18TH AND 19TH CENTURY

Abstract: This paper presents research aimed at achieving better OCR quality in large scale digitisation of newspapers and books, and opening possibilities of full-text search of digitised old Slovenian printed texts, which should enable digital library end-users to gain better transcriptions of digitised contents. The paper describes on-going work undertaken by the National and University Library of Slovenia and the Jožef Stefan Institute in the framework of the EU research project IMPACT – Improving access to text – to develop high-quality datasets, in particular ground-truth transcriptions (a clean corpus) and a lexicon of historical Slovene.

Keywords: historical printed texts, Slovenian language, page segmentation, lexicon for historical Slovene, IMPACT project

1. Introduction

In the last couple of years, mass digitisation at the National and University Library of Slovenia (NUK) and other Slovenian libraries included many historical newspapers, as well as books, which were afterwards published on the portal of the Digital Library of Slovenia – dLib.si [5, 15, 3]. The title selection was made in accordance with the overall cultural heritage digitisation strategy of NUK [20] and was focused primarily on the most important Slovenian newspapers printed from the second half of the 18th to the first half of the 20th century. The digitised collection of newspapers was composed of page scans, so full-text searches were not possible without further processing. In order to arrive at a transcription of the digitised documents, programs for scan refinement and optical character recognition (OCR) were used. Statistics shows that full-text search and quick access to contents are very important for the users of the dLib.si, because they enable them fast and easy access to transcriptions (see Chart 1).

Full-text search is possible when OCR results (i.e. plain texts) are available. On the dLib.si portal plain text is available if integrated in HTML previews or in PDF files. The statistics of dLib.si content access in the first four months of the year 2011 shows, that users use both, HTML and PDF, in 49% although HTML version is much harder to follow (see Fig. 1).

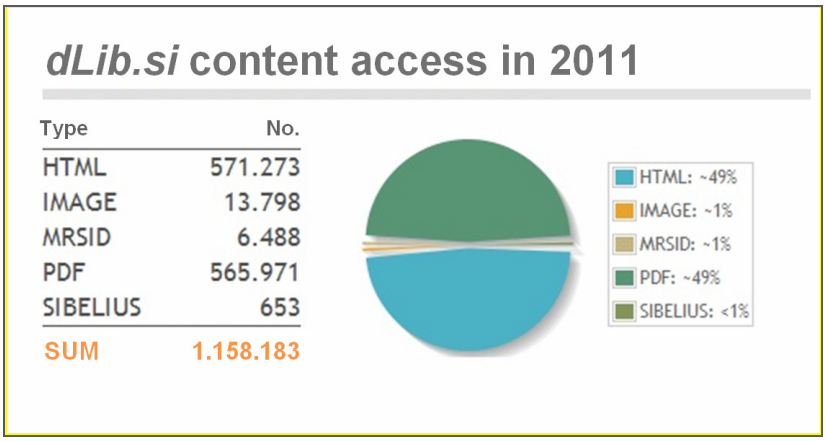

Chart 1: Statistics of the dLib.si content accesses in first four months of the year 2011

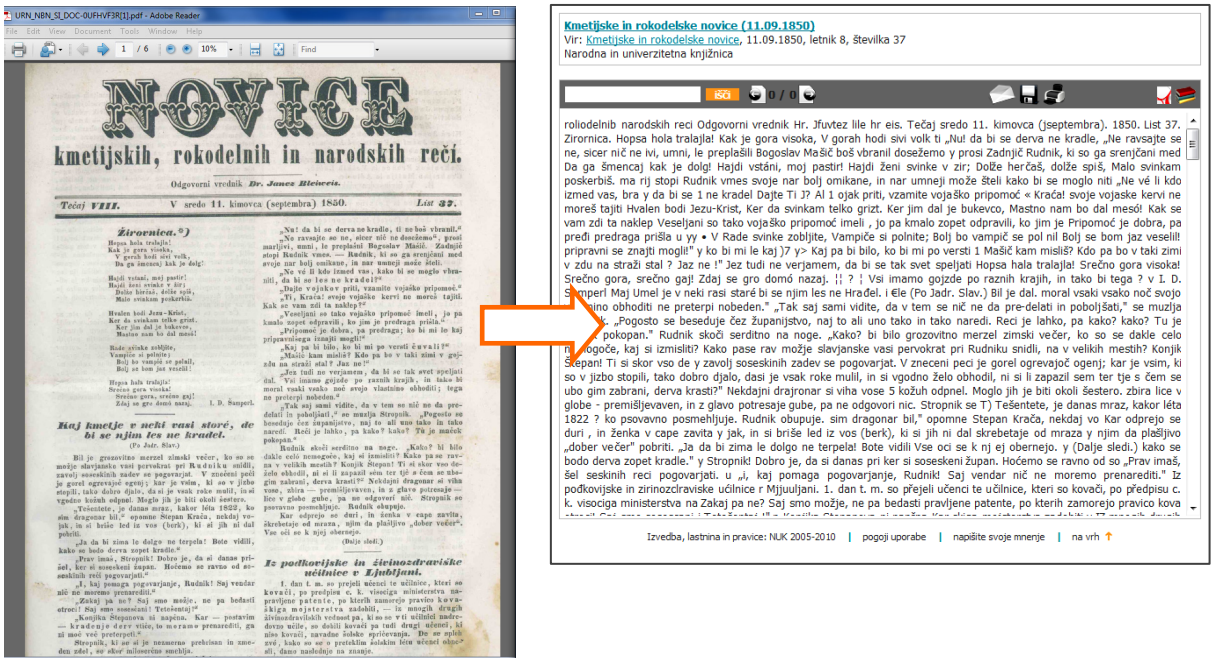

Figure 1: PDF file (left) and HTML plain text preview (right)

OCR tools were used in first historic newspaper digitisation projects from the early 1990's. After a decade, the technology had advanced and in 2008 OCR could already be used in large projects for large-scale digitisation of newspapers in the United Kingdom and Austria [12]. But nowadays state of the art is that OCR tools are mainly developed for modern languages. When applying them to historical texts the outcomes are usually full of errors and for this reason almost useless.

Problems in OCR transcriptions are caused by the language and typography from the time of the production of the originals which differ from the ones currently used. This causes major errors in the OCR and further errors occur due to the bad condition of the scanned originals (specific problems are discussed further on). The possibilities of new solutions in page segmentation and digitisation flows were analysed in the project IMPACT [13] of the 7th EU framework programme and partially also in the project EOD [6] of the EU Culture 2007-

2013 programme [14].[1] The aim of the IMPACT project is to fully integrate intellectual contents into modern information and communication technologies environments, which can only be achieved by full digitisation with transforming digital images of newspapers as well as books into electronic text.

In collaboration with Jožef Stefan Institute (JSI) we first focused on building a dataset of high-quality transcriptions of historical texts (a corpus or so called groundtruth) and, on this basis, the second step was to develop a lexicon for historical Slovene. Such a lexicon can serve to automatically correct OCR transcriptions, to enable full-text search over historical texts, and to modernise archaic language.

2. Creation of corpus

For the IMPACT Project NUK and JSI made a selection of typical documents from the 19th and the second half of the 18th century to compile a representative corpus of digitised materials. The selection of materials was primarily based on language form and the historical framework. The selection included some of the most important already digitised Slovenian historical documents of the dLib, which have poor OCR results. A further set of documents was taken from the AHLib[2] digital library, which contains Slovenian books from the period 1848-1918, which were translated from the German original. For the IMPACT corpus, a number of AHLib books, which had been already scanned, but not proofread, were selected. In addition, part of the collection of old books digitised through the service EOD was included. Since the service is aimed to deliver better text-search functionalities, the OCR in this collection has already been manually corrected.

In total, 41,313 of digitised pages of historical newspaper issues and full books from the 18th to the end of 19th century were selected for the NUK institutional dataset and afterwards incorporated into IMPACT repository[3]. A smaller subset of 5,497 scans is intended for production of groundtruth (GT), to serve as a development set for historical Slovene. Besides books, this subset also contains 123 newspaper issues, where, on average, one newspaper page has ca. 4,000 characters (cf. Table 1).

Characters per page:						
	Front page	RS I.	RS II.	RS III.	RS IV.	Average
Words	714	1190	917	1034	892	949
Characters without spaces	2500	4642	4517	4354	4253	4053
Characters with spaces	3149	5613	5528	5267	5037	4919

Table 1: Average number of characters on the front pages and on randomly selected newspaper pages; RS stands for Random Sample.

[1] IMPACT and EOD projects positively influenced the work of NUK and are evaluated as important contributors in the process of building the digital library collection in Slovenia.

[2] The AHLib library took shape as a project of the Austrian Academy of Sciences [18, 2]. The interface for manual correction and structural XML markup of AHlib is described in Erjavec [7].

[3] IMPACT repository is a common project repository hosted by Prima in which all library partners (NUK, the British Library, Biblioteca Nacional de España, Bibliothèque nationale de France, Bayerische Staatsbibliothek, Koninklijke Bibliotheek, Národní knihovna České Republiky, St. Cyril and Methodius National Library Bulgaria, Österreichische Nationalbibliothek and Poznań Supercomputing and Networking Centre for the National Library of Poland) pool their research datasets [17].

3. Production procedures

3.1 GT production. GT production was carried out by NUK in four phases:

1. *Digitisation and pre-processing of scans*. The first phase is carried out prior to the OCR process. In this phase, for example, digitised pages can be transformed from colour to black and white with binarization, to get better OCR quality and accuracy of the texts. Another useful pre-processing operation is elimination of page borders.
2. *OCR procedure*. OCR software analyses digital pages and produces text on the basis of comparisons between the identified characters and the set of pattern images stored in the software's database. OCR software also takes into consideration segmentation of every digitised page and built-in dictionaries for different languages.
3. *Post-processing of the text*. Manual correction of the OCR outcome can be applied in cases when high accuracy of the text is needed. In the past, this phase was most necessary in the case of texts published in historical languages, in black-letter or some other problematic typefaces or when the physical appearance of the digitised material highly affected the results.
4. *Evaluation*. Quality assurance can be done on several levels: on the scan and on the dataset level; [4] on institutional and on the project consortium level.[5] Once the first samples have been produced, it is advisable to do exact post-evaluation and error spotting, which ensures consistency of approach in further production.

All three phases are used in processing historical newspapers in large-scale mass digitisation. Further on in the paper we will focus on the third phase, which included editing of automatic OCR to produce transcriptions with at least a 99.95% character accuracy rate. The OCR outcomes were encoded in PAGE XML [16].

The production process of GT is represented in the Figure 2.

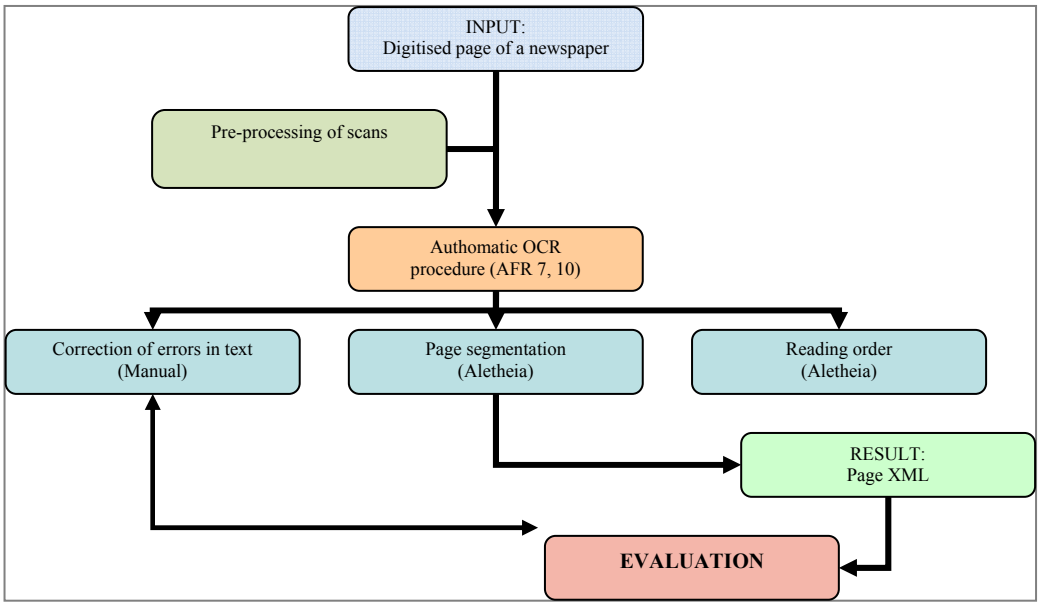

Figure 2: Graphical representation of the production of a single Page XML for GT.

[4] A special validation tool was also designed by the IMPACT project group to be used for the project needs.
[5] On the project consortium level the evaluation check was done according to the ISO 2895-1.

While producing GT, we encountered some basic problems, which affected the OCR outcome, esp. at processing newspaper pages. The following characteristics of the dataset used in this research encompass most of the problems and high risks for automatic OCR:

- two basic historical alphabets (Bohoričica / Gajica)
- Latin script and black-letter typefaces
- historical language / vocabulary
- special characters, digraphs and ligatures
- newspaper page structures / columns
- irregular spacing between letters, words and columns
- irregular / changing font sizes
- poor vocabulary recognition
- poor paper quality
- inconsistent inking
- staining (water, foxing, etc)

Further on some of the main problems are addressed. They are first of all related to bad condition of newspaper originals, most common reason for both - an inferior quality of scans and consequently also bad OCR outcomes. Especially in older editions, mechanical damages are common. A damage is caused by extensive use or because newspapers are not durable goods, printed on low quality paper and improperly stored (i.e. acid paper degradation causes foxing on paper, which then interferes with the OCR tools). Library stamps, handwritten annotations and other later interventions can also affect the OCR results. Tight back bindings make volumes difficult to open for scanning, which affects the digitisation procedure and so scans are again of low quality, which then causes major errors in the OCR outcomes. Errors which occur in the progress of scanning can be reduced to some extent with different programs in the pre-production phase. In the IMPACT project, new programs are being developed for reduction of such problems.[6]

The structure of the page or its segmentation is the second major error inducer. OCR tools have problems with the reading order or how paragraphs or articles are read one after another. Text of newspapers is usually printed in two or more columns, which also makes determining the reading order difficult. Furthermore, marginalia, headers, footers and floating text can mix with the regular texts of the articles. Graphical and image elements between the articles can affect the outcome of the OCR. For this reason it is necessary to adapt the reading order manually or in automatic settings and to indicate how the text is sequenced (see Fig. 3). For addressing this issues the IMPACT project group has developed several additional programs.[7]

Two other problems are also significant, but no proper solution has been found yet. The first problem is related to typographies and the second one to colour prints. Typographies, for example, used for printing title pages, article titles and subtitles, etc., can be of different designs. Sometimes typography is recognised in the OCR tool as an image and the text is left out, or only a part of the text is recognized and transcribed correctly (see *Figure 4*). This mostly appears in advertisements and with main newspaper titles.

[6] A set of software tools is being developed for manipulating scanned images in order to improve the recognition results of OCR engines [4, 10].

[7] *Segmentation toolkit* was developed as a set of software tools for recognising and segmenting important features of scanned documents, such as blocks, lines and characters [19]. *Functional Extension Parser* is a set of web services that can be exploited to automatically detect and tag structural metadata of scanned material has also been developed [9].

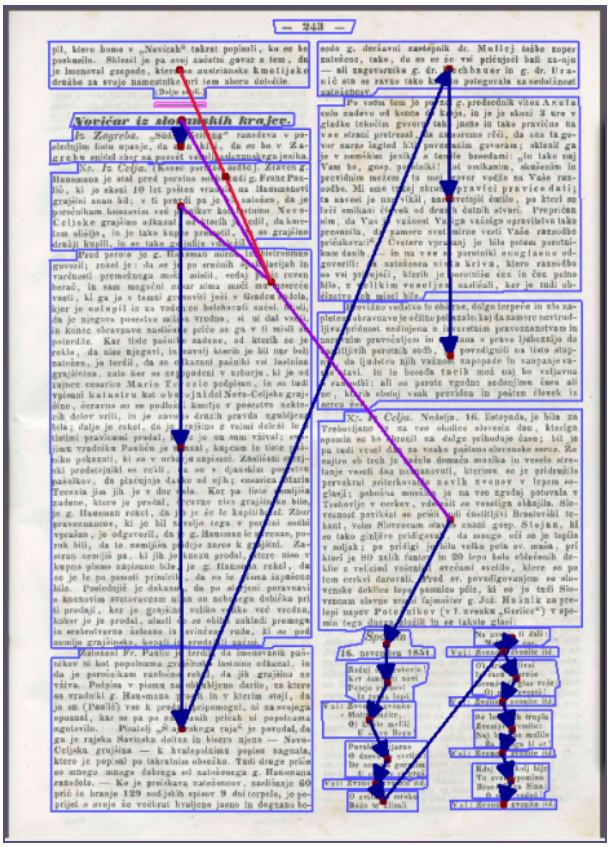

Figure 3: Reading order in the text can be set manually or with automatic settings

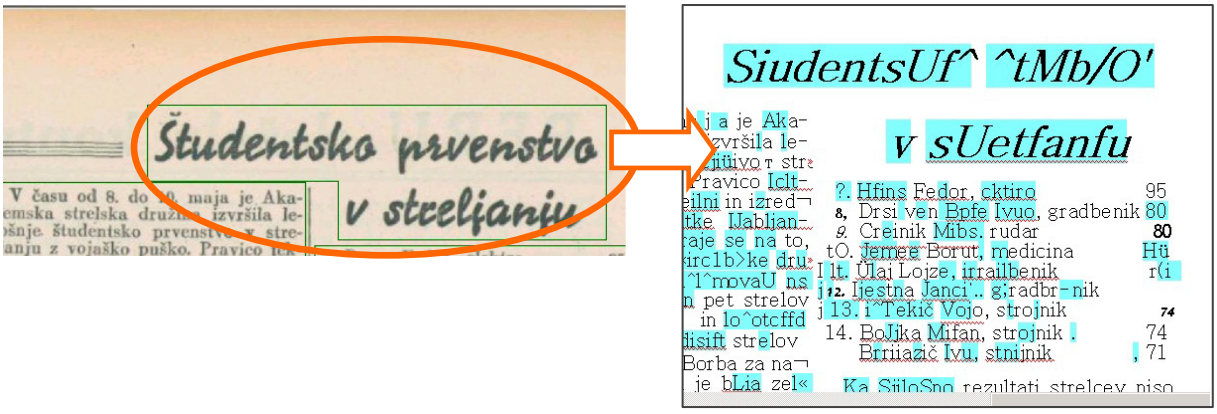

Figure 4: The title of an article is of different typography and for this reason text is badly unrecognized.

Colour print in newspapers of the 20th century can worsen the OCR results; especially when an image in dark colours is inserted for the background and the text colour is bright or white (see Figure 5). Whereas if the background is only in one colour and the contrast between background colour and text colour is not too big, then the outcome of the OCR can usually be better.

Figure 5: An example of a 20th century magazine
with white title printed on dark colour background / picture

3.2 Lexicon building for historical Slovene. Offering historical texts in the context of a digital library brings a number of problems. Firstly, as mentioned above, OCR results are much worse than for modern day texts, so the quality of text transcriptions is low. Secondly, full-text search over historical language[8] is difficult, as the texts are not lemmatised and use different orthographic conventions with different archaic spellings, typically not familiar to the user. At last, comprehension of the texts for most users can also be problematic, as texts printed before 1850 use the Bohoričica alphabet.[9]

These problems are being addressed by building a lexicon of historical Slovene language, which gives the mapping of historical word-forms into their modern-day equivalents with associated lemmas. The lexicon is being built to be incorporated in OCR tools for better recognition process, to enable modernisation and lemmatisation of historical text for full text searching, as well as to facilitate reading.

The lexicon for historical Slovene, being developed by JSI, has a simple structure. Each entry contains the following fields:

- a word-form that has been witnessed in a proof-read historical text
- the equivalent word-form from contemporary Slovene
- the contemporary lemma of the word-form
- the lexical morphosyntactic properties of the lemma.

Compiling such a lexicon with sufficient coverage is a non-trivial process: a representative corpus of proof-read historical texts must be available, a comprehensive modern-day lexicon must be obtained, and the word-stock of the former must be matched against the latter. Problems arise in methodological issues (not all historical word forms even have a modern day equivalent) as well as technological ones (having a good software environment for lexicon construction) [8].

[8] The work on newspapers concentrated on the Slovene language from the 19th century. The problems are of course, worse going further back in time, but even here, due to the late development of the written Slovene word and its spelling standardisation, there are substantial differences to contemporary Slovene.

Developing the lexicon of historical Slovene is still an on-going process, and is very time consuming activity demanding a lot of precision. To date, the lexicon contains about 3,000 words, which were selected to be the most frequent word-forms in the AHLib corpus.

3.3. Production tools. At OCR of Slovene contemporary printed texts, 90% accuracy can be achieved on the character level by using standard OCR tools (Abbyy FineReader - AFR) [1]. With the older texts and historical fonts (used for example in the 18th or the 19th century books) the quality of OCR is much lower (on average from 70 to up to 85% on the character level depending on the scan quality) and we need to invest a lot of effort in manual post correction. The average OCR quality of the documents, available on the Digital Library of Slovenia is calculated to be only 65–70% on the character level. Character uncertainty in OCR for newspapers, written in Bohoričica, is with AFR between 4.5% and 12% per page with app. 4,000 characters, unrecognized characters and other mistakes not included[10].

Different AFR versions were used for the OCR in the GT process. Some 30 test sample images were made with AFR 8 [1]. AFR 8 produced quite satisfying results on the character level for the Gajica alphabet with the binary (black and white) images, but unfortunately the results were very poor with colour images in low resolution for both Bohoričica and Gajica, where the OCR accuracy was only 58% on the character level. The text recognition was poor on the character as well as on the word level, and not even every text line was recognised. For this reason a new version, AFR 10, was used, enabling to achieve at least 70% accuracy in colour images for Bohoričica and 85% accuracy for Gajica.

Besides AFR, a special tool designed for the project needs (Aletheia) [13] was used in the production of GT. In Aletheia, the page segmentation and reading order were given and Page XML files were produced.

A web-tool LeXtractor for collaborative construction of lexica for historical language (Gotscharek et al., 2009), initially developed for German, was first adapted to the Slovene language. This involved the Slovene specific background resources, in particular a modern-day lexicon and rule set, to map historical forms to their modern day equivalents. Some internal program changes were also necessary, e.g. to enable the correct treatment of the characters of the Slovenian alphabet, and to adapt LeXtractor to the format of the Slovenian lexicon and historical corpus [13].

4. Conclusion

Modern library users are used to work with digitally born documents and expect library contents to be available online through digital libraries, such as the dLib.si. They are mostly well acquainted with nowadays technologies and expect good transcriptions of texts from the digitised materials. Their difficulties with full text search and unsuccessful searches are sometimes expressed through their requests for help and further assistance, addressed to moderators of the dLib.si. Unsuccessful searches are in most cases caused by errors in the optical character recognition (OCR).

OCR can dramatically increase the usability of digital libraries but the problems discussed in the paper can significantly impair the quality of the OCR outcomes. The main goal of the IMPACT project is to push innovation in OCR technology and language technology for historical document processing and retrieval, and to share expertise to build capacity in digitisation across Europe.

[10] The OCR accuracy is on the character level, if the OCR program recognized a character as a wrong letter this was not reported as a problem but in fact the meaning of the word could be wrong.

By cooperating in this project, NUK has produced an important set of texts with 99.95% character accuracy rate. Besides the project development needs, text of newspapers will also be used in the dLib and available to general public. Additionally, JSI has made first steps towards a lexicon of the 19th century Slovene. Linking parallels between historical words and contemporary Slovene words were established and advanced by spelling variation patterns.

We are looking forward to undertaking a third phase of the development, which unfortunately was not foreseen in the IMPACT project. It will consist of the development of a text-search tool which will enable searches in old texts but using the modern language. The background of this tool will be the lexicon developed by Jožef Stefan Institute.

All the work will finally result in a corpus of extremely usable contents, which will be available for users in the Digital library of Slovenia - dLib and will allow them to search with modern vocabulary in the database of articles written in the old Slovene language. The lexicon will be an invaluable support to the researchers and students of different fields, working with Slovene historical texts.

References

[1] *Abbyy FineReader* – program description available at http://finereader.abbyy.com/; further information available at http://www.abbyy.com/Default.aspx?DN=dddaccf1-d68f-4d53-a7dc-c1319f27e227.

[2] *AHlib books selection* available at http://nl.ijs.si/ahlib/.

[3] Ambrožič, M. [et al.]. *The Digital Library of Slovenia development strategy - dLib.si: 2007–2010*. Ljubljana: National and University Library, 2007.

[4] *Border removal tool*, information availabe at http://www.impact-project.eu/taa/tech/imgenh/bm.

[5] *Digitalna knjižnica Slovenjie – dLib.si,* available at http://www.dlib.si/.

[6] *EOD project*, information available at http://www.books2ebooks.eu/about.php5.

[7] Erjavec, T. *Architecture for Editing Complex Digital Documents.* In: Digital Information and Heritage. Zagreb : Odsjek za informacijske znanosti, Filozofski fakultet, 2007.

[8] Erjavec, T. [et al.]. *Towards a lexicon of XIXth century Slovene*. In: Proceedings of the Seventh Language Technologies Conference : proceedings of the 13th International Multiconference Information Society, volume C, 2010, http://www.impact-project.eu/fileadmin/Editorial/Documents/ISJT10_paper_LMU_and_JSI.PDF (29.4.2011).

[9] *Functional Extension Parser* information available at http://www.impact-project.eu/taa/tech/extpar/.

[10] *Geometric correction tool*, information available at http://www.impact-project.eu/taa/tech/imgenh/gcpc

[11] Gotscharek, A. [et al.]. *Enabling Information Retrieval on Historical Document Collections - the Role of Matching Procedures and Special Lexica*. In: Proceedings of the ACM SIGIR Workshop on Analytics for Noisy Unstructured Text Data, doi: 10.1145/1568296.1568309 (04.05.2011).

[12] Holly, R. *How Good Can It Get?* In: D-Lib Magazine, 2008, vol. 15, no. 3/4, doi: 10.1045/january2008-klijn (29.4.2011).

[13] *IMPACT project - Improving access to text*, information available at http://www.impact-project.eu/.

[14] Kavčič-Čolić, A. *Achieving Library Development Through European Projects: the Case of the National and University Library of Slovenia, 2010.* Presented at http://www.hkdrustvo.hr/hr/ (29.4.2011), and published in: *Slobodan pristup informacijama : 10. okrugli stol : zbornik radova*. – Zagreb : Hrvatsko knjižničarsko društvo, 2011, pp. 32–41.

[15] Krstulović, Z., Šavnik, M. *Library Material Digitisation Project Coordination and e-content National Aggregator in the Field of Culture*. Presented at SEEDI 2011 http://www.nsk.hr/seedi/ (article not published yet).

[16] Pletschacher, S. Antonacopoulos, A. *The PAGE (Page Analysis and Ground-truth Elements) Format Framework* In: 20th International Conference on Pattern Recognition (ICPR), 2010,http://www.impact-project.eu/fileadmin/Editorial/Documents/ICPR2010_The_PAGE_Format_Framework_USAL.PDF (29.4.2011).

[17] *Prima repository*, information available at http://www.prima.cse.salford.ac.uk:8080/impact-dataset/.

[18] [18] Prunč, E. *Deutsch-slowenische/kroatische Übersetzung 1848-1918 [German-Slovene/Croatian translation, 1848-1918]*. In: Ein Werkstättenbericht. Wiener Slavistisches Jahrbuch, 53 / 2007.

[19] *Segmentation toolkit*, information available at http://www.impact-project.eu/taa/tech/st/.

[20] Šavnik, M. Musek, T. Baždar, S. *Digitisation of old newspapers in the National and University Library of Slovenia (NUK) – examples of good practice*. Presented at SEEDI 2011 http://www.nsk.hr/seedi/ (article not published yet).

ines.jerele@nuk.uni-lj.si
tomaz.erjavec@ijs.si
dasa.pokorn@nuk.uni-lj.si
alenka.kavcic@nuk.uni-lj.si