

CHALLENGES IN DIGITIZATION AND OCR FOR METADATA EXTRACTION FROM THE JOURNAL ARCHIVE FOR ALBANIAN ANTIQUITY, LANGUAGE AND ETHNOLOGY

Aleksandra ANDRIĆ¹, Marija HRKALOVIĆ²

¹ Department of Library and Information Science, Faculty of Philology, University of Belgrade

² Department of Albanian Studies, Faculty of Philology, University of Belgrade

*Corresponding Author: e-mail: aleksandra.andric@fil.bg.ac.rs

Abstract

The paper presents the extraction, digitization and metadata analysis of the journal *Archive for Arbanas Antiquity, Language and Ethnology*¹, which represents a significant source for the study of Albanian cultural, linguistic, and ethnological heritage. The original issues of the journal from 1923-1926 were used as an initial resource. Python tools and libraries were used for text extraction, including PyMuPDF and Tesseract OCR, with appropriate language models for multiple languages and scripts. Metadata extraction received special care, covering title, author, year of publication, section, volume and pagination. The automatically obtained metadata was then compared with a manually created metadata database, where the similarity of titles by books was analyzed. The results showed that the OCR procedure can significantly speed up the processing of materials, but that the greatest challenges arise with multilingual titles, diacritical marks, different fonts, and longer bibliographic units. Due to the above-mentioned differences, significant deviations occur in the automatically extracted data. Bearing in mind that the journal is more than one hundred years old, recalling its importance and digitizing its basic metadata can contribute to the preservation and future reuse of the information contained in these rare issues.

Keywords: Archive, philology, multilingual material, metadata, digitization, text extraction

1. Introduction

Different kinds of old printed materials lose their readability with the passage of time, especially if not kept properly on premises. The twentieth, and especially the twenty-first century, brought with it the need for digitization in all areas. Old and rare printed material must be protected and reconstructed in some way. The *Archive* provides unique testimonies in the field of Albanian and Balkan philology, language history, folk art, history, archaeology, ethnology and numerous other topics. In chapter 2, the historical and cultural significance of this journal will be highlighted. Learning about its importance and the rarity of original printed editions motivated the authors to create a metadata database of this journal. In chapter 3, the preparation procedure for data digitization is presented in detail. Then, the process of extracting text from digitized PDF documents is shown, with special reference to the problems that arise due to multilingual content, different fonts, and the presence of diacritical marks. Special care is given to the extraction and structuring of metadata, such as title, author, year of publication, section, volume, and pagination. The automatically obtained metadata was compared with a manually created metadata base to assess the accuracy and reliability of the applied procedure. Graphical displays using heatmap and wordcloud enable a clearer comparison of similarities

¹ The journal *Archive for Arbanas Antiquity, Language and Ethnology* will hereafter be referred to as *Archive*.

<https://doi.org/10.62792/ut.filologija.v14.i26-27.p3296>

between titles and the observation of the most frequently represented terms in the metadata base.

2. Historical Context, Significance and Languages of the Papers in the Archive

The first initiative for the institutionalized study of the Albanian language at the University of Belgrade appeared in 1907. The initiative - "to study the ethnography, language and speech of the Arbanas" was initiated by the then-rector of the University of Belgrade, Jovan Cvijić, who was dedicated to the idea that "our linguists should compile a grammar and collect Arbanas folklore" (in Malja Imami, 2018: 244).² This initiative, however, bore fruit only a few decades later, when the subject History and Grammar of the Arbanas language was introduced as a part of the Seminar for Comparative Grammar of Indo-European Languages in 1920/1921. This moment is considered to be the beginning of Serbian Albanology, largely owing to the work of Professor Henrik Barić. The Seminar for Arbanas' philology was established soon after, already in 1925, and its management was entrusted to Professor Barić. At that time, the Faculty of Philosophy had only four other philological seminars - that of Serbian, Slavic, Oriental and German language (Vraneš, 2013: 528).

However, before the Seminar itself was even established, professor Barić had founded a journal - Archive for Arbanas Antiquity, Language and Ethnography (serb. *Arhiv za arbanasku starinu, jezik i etnografiju*), which was published between 1923 and 1926. Its owner and publisher was a famous Serbian ethnologist and historian, professor Tihomir R. Đorđević, while Henrik Barić was the editor.

Six volumes across three books were published between 1923 and 1926, while the entire print-run of the fourth book, which was printed in Ljubljana, disappeared under never-clarified circumstances during a period in which the relations between the Kingdom of Serbs, Croats and Slovenians and Albania began to deteriorate (Malja Imami, 2019: 131). It was only in 1969 that the Institute for Albanian studies in Prishtina printed an incomplete content of the fourth book, while publishing its first book on that occasion. That was a supplement to the reprint of the three original books of the *Archive* that were published in 1971 by the publishing house "dr Rudolf Trofenik" in Munich. Its editors were famous linguists Idriz Ajeti and Hasan Kaleshi (Kisić: 2010).

The *Archive* was a famous and internationally recognized journal of great importance. It was actually the first journal on Albanian studies in the world, in which both foreign and domestic distinguished scholars published their works (Đorđević and Barić: 1923: para. 3.; Malja Imami 2019: 130; Hrkaloć, 2022: 4). Many famous scholars, such as Holger Pedersen, Norbert Jokl, Karl Treimer, Leo Spitzer, Theodor Capidan, Rudolf Meringer, Jovan Cvijić, Stanoje Stanojević, Milan Budimir, Petar Skok and others have supported the publishing of the journal (Đorđević and Barić: 1923: para. 6).

Decades earlier linguists such as Jernej Kopitar (1829) and Franz Mikloshich (1861) had noted similarities between different Balkan languages, while Nikolai Trubetzkoy (1928) referred to relations such as those between Balkan languages as "Sprachbund". As pointed out by Tomić (2006: 1), Trubetzkoy argued that the languages of the Balkans show similarities in sentence-structure and word-formation, though they do not show systematic sound correspondences, and that they share a large number of common "cultural" words, even though their basic vocabularies may be very different. Tomić points out that the famous German linguist Gustav Weingad (1928) considered Albanian, Romanian and Bulgarian as typically Balkan languages, and Serbian, Greek and Turkish as geographically Balkan languages. Danish linguist, Kristian Sandfeld (1930), on the other hand, argued that Balkan characteristics are manifested in Greek, Bulgarian, Albanian and possibly Serbian, and that Turkish has many

² All translations from Serbian sources are my own.

lexical concordances with each of them individually. Many other linguists have since studied the Balkan Sprachbund, and it continues to attract interdisciplinary attention of scientists to this day. The late 19th and early 20th century marked a period that witnessed a growing scientific interest in the Balkans and Balkan languages, eventually laying the foundations of Balkanology as a scientific discipline. It is easy to see why a journal such as the *Archive* was founded in these historical circumstances and why it attracted so much international attention. Ever since Kopitar, it was evident that the Albanian language plays an important role in the Balkan Sprachbund, and the fact that the *Archive*, in its very title combined the Albanian language, ethnology and antiquity provides a clear indication that the analysis of the Balkan languages cannot be restricted to grammar alone. The historical coexistence, customs and culture of the region's speakers are an important part of its very nature.

As the editor himself, Henrik Barić, and the journal's owner, Tihomir Đorđević, acknowledge in the foreword to the first volume, the *Archive* had a difficult task to introduce the original works, contributions and material from all fields of Albanology to the public, to provide critical notes alongside a systematic bibliography of all scientific works in history, language, literature, archeology, political history, folklore, anthropology and mutual relations of the Albanians and their neighbors, as well as to provide the most complete picture as possible of the work that has been done in the field of Albanology "at home and abroad" (Đorđević and Barić, 1923: para. 5-6). Đorđević and Barić pointed to the emergence, at that time, of stronger interests to examine the past, customs and language of our old neighbors, Albanians. They point out the following:

This journal, which is also the very first Albanological journal, was launched in response to this movement and to the obvious need to organize and consolidate, immediately, from the beginning, the closest possible cooperation of all our scholars whose research, whether folkloric or linguistic, historic or archeological, falls within the field of the Albanian studies. Therefore, we hope that by establishing the "Archive" we will satisfy the scientific need that has been felt in academic circles both within and beyond our country's borders. Accordingly, the works will be published not only in Serbo-Croatian and Slovenian, but also in French, German and English as well. (Đorđević and Barić, 1923: para. 3) ³

This alone shows us that the *Archive* would publish its works in several world languages - in addition to the official languages of the Kingdom of SHS, it would publish in languages that were considered as the pillars of world science at the time. In that way, the *Archive* would open its doors to foreign authors as well and become accessible beyond the borders of its country. In the very first volume, many distinguished scholars publish their works: Petar Skok, a renowned Croatian linguist, onomastician and literary historian, analyzes toponymes believed to be of Illyrian origin; Norbert Jokl, Austrian linguist and father of Albanology; Gligoriје Elezović, Serbian philologist and orientalist, who in this volume brought "an Arnaut variant about the Battle of Kosovo"; Rajko Nahtigal, Slovenian linguist who, although not usually associated with the Albanian studies, travelled in 1917 to Albania, after being invited by dr. Gjergj Pekmezi, president of the Albanian Literary Commission, for consultation on language standardization (Hamp, 1972: 1634). In the *Archive*, he published a paper on the Elbasan alphabet. Tihomir Đorđević and Henrik Barić themselves, along with other authors, published their works in this issue. Other well-known authors published their works in subsequent volumes: Wilhelm Meyer-Lübke, the renowned Swiss linguist; the Italian linguists Carlo Tagliavini and Alfredo Trombetti; as well as Milan Šufflay, Croatian historian and politician.

³ Original text: Према овом покрету и према очевидној потреби да се одмах из почетка организује и учврсти што ужа сарадња свих оних научних радника наших, чија испитивања, фолклорска или језичка, историска или археолошка залазе у област арбанаских студија, покренут је овај часопис, који је уједно и први албанолошки часопис уопште. Ми се зато надамо да ћемо оснивањем „Архива“ задовољити једну научну потребу која се у стручним круговима осећала и ван граница наше земље. Саобразно томе радови ће излазити сем на српскохрватском и словеначком, и на француском, немачком и енглеском језику. (Đorđević, Barić, 1923: para. 3)

Moreover, the very first volume of the *Archive* contained texts written in Serbo-Croatian (latin alphabet), Serbian Cyrillic, German, Slovenian and Czech, while the titles of the papers in the table of contents were written, among others, in Latin, Romanian and Italian language and alphabet. Furthermore, the names of the authors and the titles of their papers were often written in different languages and alphabets (e.g.: *X. Bapuħ: Norbert Jokl, Albanisch y Geschichte der Indogermanischen Sprachwissenschaft seit ihrer Bedründung durch Franz Bopp* (*Archive*, 1923: table of contents)). All of this tells us that the texts published in the *Archive* were written in even more languages than was originally intended (Serbo-Croatian, Slovenian, French, German and English, as announced in the foreword by Đorđević and Barić (1923: para. 3) and in several different alphabets, which beyond doubt represents a great challenge in the process of digitization, optical character recognition and extraction of the metadata from the journal. The process will be discussed in detail in the paragraphs below.

3. Digitization and Metadata Extraction Process

This chapter presents the process of digitizing the printed edition of the journal the *Archive*, published between 1923 and 1926. Given that this journal belongs to the group of journals over 100 years old, there is a risk that a significant part of the cultural and scientific heritage contained in it will be lost over time or become difficult to access for researchers. In the digitization process, it is first assessed whether the printed material has physical damage, missing pages, worn parts or other defects that could make further processing difficult. The *Archive* has a well-preserved appearance, without major defects, although the paper and cover are showing the signs of aging.

The digitization process begins with scanning the contents of this journal. There are a number of tools that enable accurate conversion of printed text into digital format, including professional scanners, mobile apps and other document capture technologies. In the process of digitizing this publication, a Google Drive scanner was used, which enabled correct scanning quality. This results in a digital form of the resource, but we still cannot consider such content to be part of a digital archive. The material that has been scanned should be examined in detail in order to notice possible irregularities in the form of blurred pages, distorted margins, cut text, or some other omissions. After these inspections, further processing can be continued.

Content saved in PDF format can be extracted using libraries for processing and reading PDF content or using OCR tools. Depending on the language and script used, as well as the amount of images, tables and other objects that are present, the processing process differs. The *Archive* brings together multilingual content. These linguistic diversities make the text extraction process difficult. Other factors that hinder content processing are tables, which do not have precisely bounded columns and rows. Images are not represented, and they are otherwise considered the most difficult object to extract.

For extracting text from digitized PDF documents, the PyMuPDF⁴ library was used for processing PDF files and reading the textual layer, while OCR processing had previously been performed using the OCRmyPDF tool, which relies on the Tesseract OCR system⁵. Appropriate language models were applied for English (eng), Serbian in Cyrillic script (srp), Serbian in Latin script (srp_latn), Albanian (sqi), Greek (ell), Turkish (tur), Slovenian (slv), French (fra), German (deu), and Italian (ita). It is expected that there will be no 1-1 match of the text before and after the extraction, but the results are still of satisfactory quality. Full text matching requires manual processing, which is most often used when creating the gold standard for machine learning models.

There are two approaches for further processing: either structuring the entire content

⁴ PyMuPDF. (n.d.). PyMuPDF documentation /on-line/. Retrieved January 26, 2026 from: <https://pymupdf.readthedocs.io/en/latest/>

⁵ Tesseract OCR. (n.d.). Tesseract OCR: Open Source OCR Engine /on-line/. Retrieved January 26, 2026 from: <https://github.com/tesseract-ocr/tesseract>

using known formats such as XML, TEI recommendations, JSON and/or extracting metadata that will be then saved in one of the formats that is still searchable. Given the numerous challenges previously mentioned, this paper deals only with journal metadata, not the actual full content of the papers in journal. To extract the metadata, we observed the page numbers of the papers in the journal. For title, author, pagination, and section information, refer to the journal table of contents. Headers also contain title and author information; however, these are written alternately on nearly every page of the article and are often in a small font that is difficult for OCR to work properly. The main goal of further processing is the transformation of the table of contents into a structured set of data suitable for further analysis and integration into metadata databases. As a result, a record was created for each content item with the following fields: book, volume, year of publication, section, author, title and pagination.

The extraction was achieved in the Python programming language using the PyMuPDF library, where instead of the classic "pure text" extraction, an approach based on the spatial coordinates of words was applied. Individual words are obtained along with their coordinates on the page, which allows the reconstruction of rows and columns even when OCR violates the original layout. Based on the y-coordinates, the words are grouped into rows. Then, for each line, the left text part (author/title) and the right part containing pagination were separated. Pagination is recognized by regular expressions that support different hyphen types and range shapes (e.g. 133-134), as well as single pages (e.g. 167). In cases where only the starting page is listed in the table of contents, the end of the range is estimated based on the next item, considering possible overlaps when the next item starts on the same page. Section identification is performed by strict comparison with a predefined set of expected section names. To mitigate OCR deviations in individual letters, text normalization and "fuzzy matching" (approximate comparison) were applied. This reduces the risk of parts of the title being misinterpreted as section names.

In the end, the author and title were separated from each item by applying the rule of separation by the first colon, i.e. by the form "Author: Title", while records without a colon were treated as cases in which the title appears without an explicitly mentioned author in the same line. All extracted data were saved in JSON format. The metadata obtained in this way does not completely match the content in the PDF, but the extraction process is still incomparably faster than completely manual processing. The goal is to compare the automatically obtained data with the metadata database that was manually created for the purposes of this work. This will give us an evaluation of the accuracy and reliability of the automatically extracted metadata compared to the manually created database. Although there is a possibility that certain errors may occur in the manually formed database during entry, we will consider them to be negligible. All scripts and languages that appear as metadata values are taken into account.

Table 1 presents the layout of a part of the manually created database and the corresponding part of the database obtained by automatic extraction. The values for volume, issue and year were manually predefined and inserted as fixed fields during the automatic extraction process.

Table 1

Appearance of a part of the manually created database and automatically extracted data.

<pre>{ "volume": "1", "issue": "1-2", "section": "Критика", "title": "Acta et diplomata res Albaniae mediae aetatis illustrantis", "author": "Ст. Станојевић", "year": "1923", "pagination": "235-236" }</pre>	<pre>{ "volume": "1", "issue": "1-2", "section": "Критика", "title": "Acta et diplomata res Albaniae mediae aetatis illustrantia >", "author": "Сид. Станојевић", "year": "1923", "pagination": "235-236" }</pre>
--	--

It is easy to notice that there are certain discrepancies, and in the next chapter, the procedure for obtaining the similarity value of these two bases will be shown.

4. Evaluation and Visualization of Extracted Metadata

In this section, the titles extracted by Tesseract OCR are examined in order to identify the main challenges of processing multilingual and multiscript material. The manually created metadata database was used as a reference for detecting deviations in the automatically extracted data. The analysis was performed separately for each book. Since titles are highly informative but also sensitive to OCR errors, the focus was placed on the title field. The results highlight the difficulties caused by diacritics, different scripts, foreign-language titles and complex bibliographic entries. Before comparison, titles were normalized: Cyrillic was converted to Latin, diacritics were removed, text was converted to lowercase, and redundant punctuation and multiple spaces were removed.

A flexible approach based on multiple comparisons was used to calculate similarity: basic character comparison using the SequenceMatcher method, sorted token comparison, token set comparison, and partial title matching. The highest obtained value was taken as the final similarity value, thus mitigating the impact of OCR errors, changed word order and incompletely recognized titles. The results show that the best average similarity is achieved for the first book, where it is 76.97%, while the second book has the lowest average similarity of 60.18%, which indicates a higher number of OCR deviations or unsuccessfully recognized titles. The third book has an average similarity of 65.44%, but it has seven additional records in the automatic database, which indicates that the OCR procedure probably incorrectly separated some parts of the content as separate bibliographic units. Overall, the results show that automatic extraction can significantly speed up metadata processing, but due to errors in title recognition and structuring, it still requires manual checking and correction.

Table 2

Determining the efficiency of the Tesseract OCR for title extraction.

Book	Manua l records	Automatic records	Matched records	Missing automatic database	in Extra automatic database	in Average title similarity
Book 1	27	25	25	2	0	76.97%
Book 2	36	33	33	3	0	60.18%
Book	22	29	22	0	7	65.44%

Book	Manual records	Automatic records	Matched records	Missing automatic database	in Extra automatic database	in Average title similarity
3						

Based on the manually created database, it can be concluded that the most critical places for automatic recognition were those where there are multilingual titles, different letters, diacritical marks, and non-standard typographical forms. Titles in German, French, Italian, Slovenian and Albanian are especially problematic, as they contain characters such as ü, ä, è, é, à, â, î, ć, č, š, ž, as well as various apostrophes, hyphens and joined word forms. For example, titles such as Zur Rückerschliessung der illyrischen Götterwelt, Maximilian Lambertz, Albanische Märchen, Draske private-legal documents in the 13th century, À propos du nasalisme et du rhotacisme roumainoalbanais, L'arumeno râmân, râmân è influenzato..., as well as Beiträge zur Biographie... des türkischen Dichters Sâbit appear in the database. Such examples show that the Tesseract not only had a problem with reading individual letters, but also with preserving the correct form of words, accents, foreign graphemes and punctuation. An additional problem is the very long titles, especially in the "Criticism" and "Notes" sections, where one entry often contains several bibliographic units, author names, titles, brackets and hyphens. Previous research has shown that OCR errors are not neutral and that they can significantly affect the results of quantitative analysis of historical texts, especially when the material contains old typographic forms, ligatures and specific graphemes (Hill, Hengchen, 2019:840). Because of this, OCR may incorrectly merge or separate parts of the text, resulting in a poorer match with the manually created database. It can be concluded that shorter and simpler titles were most reliably recognized, while the biggest deviations occurred with titles with diacritical marks, a mixture of Cyrillic and Latin, older orthography, and long bibliographic strings.

Table 3
Determining the efficiency of Tesseract OCR for title extraction - flexible version.

Book	Manual records	Automatic records	Matched records	Missing automatic database	in Extra automatic database	in Average title similarity
Book 1	27	25	26	1	2	91.42%
Book 2	36	33	34	2	12	85.04%
Book 3	22	29	22	0	16	79.05%

In a more flexible version of the code, a laxer comparison of titles has been added to mitigate errors generated during OCR extraction. Unlike the original approach, one automatic record can now be the best candidate for multiple manual titles, which is important when Tesseract merges multiple bibliographic units into one longer text. In addition to the basic character comparison, additional similarity measures are introduced, such as token set comparison, token content comparison, and comparison using character n-grams. The processing of frequent OCR substitutions and foreign graphemes, as well as the removal of stop words, diacritics and non-standard characters, has also been added. Thanks to these changes, the comparison better recognizes partially recognized, abbreviated or merged titles, so higher and more realistic average similarity values are obtained.

Word cloud for titles.



Figure 1 shows a word cloud generated based on the title from the manually created metadata database `metadata.json`. The size of the words in the image depends on the frequency of their appearance in the titles, so those words that appear more often in the entire data set are shown larger. The Python programming language, the `wordcloud`⁶ library for generating word clouds, `matplotlib` for displaying and saving the image, as well as pre-cleaning the text by removing punctuation, numbers and frequent non-informative words were used to create the image.

To further analyze titles' similarity in the manually created metadata database, a program was implemented that enables the comparison of the selected title with all other titles in the database. The title "Illyro-Thrakisches" was used as an input example, after which the program found the closest matching record from the metadata.json file and calculated its degree of similarity with the remaining titles. In the processing, the Python libraries difflib⁷ were used for calculating the similarity of text strings, as well as numpy and matplotlib⁸ for creating and displaying heatmap visualization. Prior to comparison, titles were normalized by transliterating Cyrillic to Latin, removing diacritics, converting text to lowercase, and removing redundant punctuation and spaces. Similarity was calculated by combining several approaches: character comparison, sorted tokens, token sets, and partial matching, with the highest obtained similarity being taken as the final value. In the visual display, only titles with a similarity of at least 60% were retained, which highlighted more relevant results and enabled a clearer observation of thematically or lexically close titles within the database.

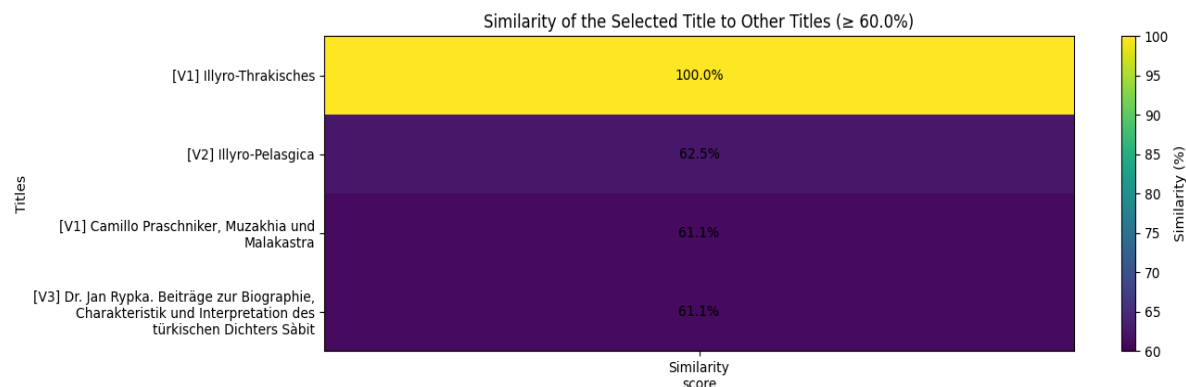
⁶ Mueller, A. (2020). *Command Line Interface*. wordcloud 1.8.1 documentation /on-line/. Retrieved May 28, 2026 from: https://amueller.github.io/word_cloud/cli.html

⁷Python Software Foundation. (n.d.). `difflib` — Helpers for computing deltas /on-line/. Python 3 documentation. Retrieved May 3, 2026 from: <https://docs.python.org/3/library/difflib.html>

⁸ Matplotlib. (n.d.). Matplotlib documentation /on-line/. Retrieved May 3, 2026 from: <https://matplotlib.org/stable/index.html>

Figure 2

Heatmap visualization of the similarity between the selected title Illyro-Thrakisches and other titles from the manually created metadata database.



5. Conclusions

Bearing in mind the historical importance of the *Archive*, we undertook the digitization of its metadata. A manually created metadata database was compared with the database obtained through automatic extraction from documents processed using Tesseract OCR. The results show that automatic metadata extraction produced a usable initial dataset, although the extracted data did not fully correspond to the manually created database. Such deviations were expected, considering the age of the printed edition, its multilingual content, the use of different scripts and the presence of foreign graphemes. The greatest discrepancies were observed in longer titles and in cases where the OCR process merged several bibliographic units into one record or incorrectly separated a single unit. A more flexible similarity-based comparison method produced better results, as it allowed partially extracted, shortened or merged titles to be recognized more accurately.

In future work, it would be useful to expand the metadata base with additional fields, improve the OCR processing procedure by using more precise language models, and perform more detailed manual correction of automatically obtained records. Also, it is possible to develop a more complete digital journal database, which would include not only metadata but also the entire text of the articles.

References

- [1]. Bondžić, D. (2013). Henrik Barić – prilog za biografiju [Henrik Barić – a contribution to a biography]. In Vraneš, A. (Ed.) *Nastava albanskog jezika kao stranog* [Teaching Albanian as a foreign language] (pp. 57–73). Filološki fakultet.
- [2]. Đorđević, T. R., & Barić, H. (1923). Predgovor [Preface]. *Arhiv za arbanasku starinu, jezik i etnologiju*, 1(1–2).
- [3]. Hamp, E. P. (1972). Albanian. In T. A. Sebeok (Ed.), *Current trends in linguistics: Vol. 9. Linguistics in Western Europe* (pp. 1626–1692). Mouton. https://www.degruyterbrill.com/document/doi/10.1515/9783_111684970/html
- [4]. Hrkalo, M. (2022). Prilog biografiji Henrika Barića [A contribution to the biography of Henrik Barić]. In *Monografija povodom 101. godišnjice učenja albanskog jezika na Univerzitetu u Beogradu = Monografi me rastin e 101 vjetorit të mësimdhënies së gjuhës shqipe në Universitetin e Beogradit* [Monograph on the occasion of the 101st anniversary of teaching Albanian language at the University of Belgrade] (pp. 39–58). Filološki fakultet Univerziteta u Beogradu.
- [5]. Kisić, M. (2010). Arhiv za arbanasku starinu, jezik i etnologiju. In *Srpska enciklopedija* [Serbian encyclopedia] (Vol. 1, Book 1). Matica srpska; Srpska akademija nauka i umetnosti; Zavod za udžbenike. <https://srpskaenciklopedija.rs/books/slovo-a/page/arxiv-za-arbanasku-starinu-jezik-i-etnologiju>

- [6]. Malja Imami, N. (2019). *Uvod u albanologiju* [Introduction to albanology] (pp. 130–134). Filološki fakultet Univerziteta u Beogradu.
- [7]. Malja-Imami, N. R. (2018). Pozitivna percepcija o Albancima [A positive perception of Albanians]. *Nasleđe: Časopis za književnost, jezik, umetnost i kulturu*, (41), 237–251.
- [8]. Pudić, I. (1956). In Memoriam Henriku Bariću [In memoriam of Henrik Barić]. *Godišnjak Centra za balkanološka ispitivanja*, 1, 1–6. <https://www.ceeol.com/search/article-detail?id=746647>
- [9]. Tomić, O. M. (2006). Introduction. In O. M. Tomić (Ed.), *Balkan Sprachbund morpho-syntactic features* (pp. 1–33). Springer. https://doi.org/10.1007/1-4020-4488-7_1
- [10]. Vraneš, A. (2013). Nastava albanskog jezika i književnosti na Univerzitetu u Beogradu [Teaching of Albanian language and literature at the University of Belgrade]. In Vraneš, A. (Ed.) *Nastava albanskog jezika kao stranog* [Teaching Albanian as a foreign language] (pp. 51–55). Filološki fakultet
- [11]. Hill, M. J., & Hengchen, S. (2019). Quantifying the impact of dirty OCR on historical text analysis: Eighteenth Century Collections Online as a case study. *Digital Scholarship in the Humanities*, 34(4), 825–843. <https://doi.org/10.1093/lc/fqz024>
- [12]. PyMuPDF. (n.d.). *PyMuPDF documentation* /on-line/. Retrieved January 26, 2026 from: <https://pymupdf.readthedocs.io/en/latest/>
- [13]. Tesseract OCR. (n.d.). *Tesseract OCR: Open Source OCR Engine* /on-line/. Retrieved January 26, 2026 from: <https://github.com/tesseract-ocr/tesseract>
- [14]. Mueller, A. (2020). *Command Line Interface. wordcloud 1.8.1 documentation* /on-line/. Retrieved May 28, 2026 from: https://amueller.github.io/word_cloud/cli.html
- [15]. Python Software Foundation. (n.d.). *difflib — Helpers for computing deltas* /on-line/. Python 3 documentation. Retrieved May 3, 2026 from: <https://docs.python.org/3/library/difflib.html>
- [16]. Matplotlib. (n.d.). *Matplotlib documentation* /on-line/. Retrieved May 3, 2026 from: <https://matplotlib.org/stable/index.html>