Automating Historical Research Processes on Romanian Texts: A Case Study on University Annals from the Interwar Period

Flavia-Elena Hațiegan[*], **Luminița Dumănescu**[**]

[*]*Babeş-Bolyai University, Doctoral School "Population Studies and History of Minorities", Cluj-Napoca, Romania, flavia.hatiegan@ubbcluj.ro*
[**]*Babeş-Bolyai University, Centre for Population Studies, Cluj-Napoca, Romania, luminita.dumanescu@ubbcluj.ro*

Abstract: This study explores the role of academic discourse in constructing national and racial identity in interwar Romania, focusing on the Annals of the University of Cluj between 1919 and 1942. By employing text analysis methods from the field of digital humanities, such as natural language processing (NLP), bigram frequency analysis, and network visualisation, we examine how nationalist and racial categories were embedded in academic speech. The research reveals the systematic integration of concepts such as race, eugenics, and national identity across disciplines, from hygiene and ethnography to philosophy and psychology. These findings highlight the university's central role in the Romanianisation process and the exclusion of ethnic minorities, particularly in the aftermath of the 1918 unification. The results also underscore the deep interconnection between intellectual production and state ideology during this formative period. While the analysis is limited by challenges in OCR quality and text standardisation, it demonstrates the value of digital tools for uncovering discursive patterns in historical sources. This interdisciplinary approach offers new pathways for understanding the socio-political functions of academic institutions and contributes to broader debates on nationalism, race, and memory in Central and Eastern Europe.

Keywords: nationalism, race, digital humanities, text networks, NLP

https://doi.org/10.24193/RJPS.2025.1.02

1. Introduction

Romanian universities during the interwar period were strongly influenced by the nationalist plan. This period in Romania's history was generally characterised by fascist movements, racist beliefs, and royal and military dictatorships (Cârstocea 2014). These factors led to violent revolts, massacres of the Jewish population, and, ultimately, to the deportation of Romanian Jews and Roma, in what is known as the Romanian Holocaust. Even today, many Romanians regard the authors of the Holocaust as national heroes—such as Marshal Ion Antonescu, responsible for the Holocaust in Romania, and members of the fascist movement, such as the leader of the Iron Guard, Corneliu Zelea Codreanu (Dan 2018: 101–102).

Although there is existing research on the fascist movement in Romania—particularly the Iron Guard (Ioanid 1990; Cârstocea 2014; 2017; Clark 2015)—on the interwar eugenics movement (Bucur 2002; Turda 2008; 2015; 2016), on the Holocaust of the Jews in Romania, and currently there is a growing interest in the study of the Holocaust of the Roma in Romania (Matei 2022; Turda and Furtuna 2022; Furtună 2018), the social dynamics of the interwar period, during which racial and hate-inciting discourses against ethnic minorities were developed, remain considerably underexplored. Many Romanian university professors and intellectuals of the interwar period held strong nationalist views that led to ideas of ethnic superiority, racism, or eugenics, which were disseminated in the press and in their public writings (Craioveanu 2020). At the University of Cluj, there was an important eugenics centre that influenced the discourse of professors from various departments unrelated to medical studies, such as the departments of Law, History, or Literature (Craioveanu 2020).

This study aims to analyse the official discourse of intellectuals from the interwar period at one of the most important institutions in Romania at the time – the University. It examines official speeches at the University of Cluj between 1919 and 1942, using new methods of automated analysis on historical texts in the Romanian language. More specifically, automated text analysis is used to uncover socio-cultural patterns related to national identity and ethnic relations in the official drhetoric of the University.

The study adapts methods used in the social sciences to historical analysis, employing natural language processing (NLP) tools, generating text networks with Python, and using bigrams to explore the relationships between the terms and concepts present in the discourse of university intellectuals. The aim of this work is to examine what kinds of ideas were used in connection with the concept of *românitate* or Romanian identity. A more thorough exploration of the official discourse used at the University of Cluj can help us better understand the dynamics of the nation-building process during the interwar period, particularly in relation to ethnic minorities, hate speech, racism, and fascism.

The analysis activity was divided into three main stages (see Figure 1). The first was data preparation, which involved cleaning the material of common errors generated during text preprocessing with optical character recognition (OCR) using Regular Expressions, and correcting misspelled words by running a spellchecker on the text, using the Language Tool library in Python. The second stage consisted of text analysis, which first involved tokenisation, or dividing the text into individual units (in this case, words), focusing only on certain parts of speech, such as nouns and adjectives. Then, the text was lemmatised, meaning that all the tokenised words were converted into their dictionary form. Next, we proceeded to identify bigrams, or pairs of words that frequently appear together in the corpus, related to terms targeted by this research, using the Gensim library in Python. More precisely, the terms central to this analysis were *rasă* (race), *superioritate* (superiority), *român* (Romanian), *străin* (foreigner), *etnie* (ethnicity), *evreu* (Jew), *țigan* (Gypsy), and *ereditate* (heredity). The bigrams were exported into a CSV file, which was then manually reviewed in order to normalise the list of bigrams and check for any outliers. Finally, in the third stage, a graphical visualisation of the extracted bigrams was generated to observe how they were interconnected, using the *NetworkX* and *Pyvis* libraries in Python.

Text analysis using Natural Language Processing (NLP), which is a component of Machine Learning (ML), itself a subfield of Artificial Intelligence (AI), has numerous applications in historical research and can yield highly accurate results. A considerable number of studies have been published in the fields of social sciences, history, and literature using NLP, demonstrating the usefulness of automated text analysis methods in academic research. While automated text analysis using NLP has been applied in Romanian literary studies, its use on historical Romanian texts remains scarce. This study represents one of the first attempts to apply NLP methods to official interwar university texts from Romania, aiming to uncover socio-cultural patterns related to national identity and ethnic relations. By employing Python-based tools such as bigram analysis and network visualisation, the research contributes to the digital humanities field by offering new methods for gaining insights into the construction of discourse in this important historical period.

In Romania, a study published in 2020 demonstrated how Romanian literary trends have historically transformed through automated topic modelling using NLP. The study analysed thematic changes that occurred over time as well as the co-occurrence of certain concepts (Neagu et al. 2020). Another study, from 2016, traces the chronological evolution of the Romanian linguistic style used in the press in Soviet-era Bessarabia and post-Soviet Bessarabia, comparing it with the linguistic style employed in Romania during the same periods (Gifu et al. 2016).

At the international level, similar methods have been used to demonstrate the effectiveness of these tools in historical and social studies, in order to "answer research questions about depictions of historically marginalized groups that have been previously studied [...] using traditional methods," (Lucy et al. 2020) yielding accurate results. These methods have also proven effective in psychology studies, helping to better understand people and culture through the examination of language (Berger and Packard 2022).

Another study published in 2020 explores the dynamic changes of key concepts used during the socialist regime in Hungary within public discourse, employing NLP methods (Szabó et al. 2020). A study from the same year, applying NLP techniques to the language used in history textbooks in Texas between 2015 and 2017, identified how ethnic, racial, and gender groups are represented in school education (Lucy et al. 2020).

Thus, NLP methods for analysing historical texts have been predominantly used at the international level and it has been proven that they can provide accurate and relevant results. In Romania, this approach to historical texts is relatively recent and certainly requires further experimentation to identify and adapt the most suitable methods to the specific demands of analysing Romanian texts.

2. Historical Context

The idea of race as a distinguishing criterion between human groups began to take shape at the end of the 16th century, when in the Anglo-Saxon world the word *race* had a looser meaning, referring to *type* or *kind*, (Smedley and Smedley 2005, 19) and the term was used in various contexts. By the 18th century, with the expansion of English colonies, it had acquired a hierarchical purpose—used to classify groups of people in colonised territories. Always referring to a power relationship between one group and another, or to the notion of cultural superiority, the concept of race as it developed in European thought placed the West—or more precisely, the white man—at the top of the human hierarchy (Turda and Balogun 2023: 2)

Historically, conceptions of human races have focused on the differences between various human groups. Only relatively recently—after the discovery that humans share 99.9% of their genetic material and that genetic differences between groups amount to just 0.01%—have racial theories begun to be challenged (Smedley and Smedley 2005: 19). In the 20th century, the dominant racial theory in the Western world combined biological elements with behavioral traits observed in humans. However, this theory has no scientific basis and is now considered by experts to be pseudoscience or a cultural construct (Smedley and Smedley 2005). Nevertheless, especially during the first half of the 20th century, racial theory remained dominant in both European and North-American contexts. The idea of race took root not only in colonial settings, but also in post-World War I Romania and across Eastern Europe, where it became present in public discourse and embedded in the collective mindset.

In Romania, the "scientific" or biological racialisation of individuals from various ethnic groups began as early as the mid-19th century, when a significant group of physicians and naturalists had already started employing notions of racial classification and undertook efforts to popularise the concept of scientific racism, following the Enlightenment tradition of classifying human life. A highly important study from 2022 offers an overview of the process of popularising scientific racism in Romania since the mid-19th century (Koszor-Codrea 2022: 37–56). In addition to its critical analysis of the 19th-century scientific discourse, the study reveals that the promoters of scientific racism were influenced by the German tradition of *Naturphilosophie* (Koszor-Codrea 2022: 43), which essentially applied organic categories derived from nature to the human mind, including society, culture, and other forms of human organisation. Moreover, the study explains how, in line with the European Enlightenment tradition of ordering and hierarchising living beings through racial classification, Romanians were placed above the Roma, who until 1856 had been enslaved in Moldavia and Wallachia for half a millennium (Koszor-Codrea 2022: 37–56).

The formation of Greater Romania (România Mare) in 1918, following the union of the Kingdom of Romania with Transylvania, Banat, Bukovina, Bessarabia, and earlier with Dobruja, marked a significant shift in the country's political and cultural landscape. These regions had previously been under the control of the Austro-Hungarian, Tsarist, and Ottoman Empires. The creation of the Romanian state began in 1859 with the election of Alexandru I. Cuza as ruler of the Principalities of Moldova and Wallachia, which had been separate entities until then. The unification of the two principalities was formalised in

1862, when a government was formed in Bucharest, the new capital. After the Russo-Turkish War of 1877–1878, also known as Romania's War of Independence, Romania gained its independence from the Ottoman Empire and, in 1881, became the Kingdom of Romania under the rule of King Carol I of the German Hohenzollern-Sigmaringen dynasty.

The period following the First World War brought about a series of fundamental transformations in Transylvania, from a social, institutional, political, economic, and demographic perspective. The war had severe demographic consequences for the country, and mortality remained high in the post-war years, especially due to widespread disease. The unification of Bessarabia, Bukovina, and Transylvania with the Kingdom of Romania in 1918 resulted in the doubling of Romanian territory and a tripling of the population of the newly formed Romanian state. At that time, Romanian society was still predominantly agrarian, with 80% of the population living in rural areas (Livezeanu 1995: 9). The unification of the principalities also led to a significant increase in the proportion of ethnic minorities, reaching 30%, most of whom lived in urban centres (Livezeanu 1995: 9). In post-war Transylvania, the contrast between the rural majority and the urban minorities resulted in what has been termed an ethnic confrontation (Livezeanu 1995: 11). Given that interwar Romania was a newly established state undergoing a process of modernisation and national consolidation, nationalism—somewhat organically—became a foundational element in the governance of the country and in the internal integration of its constituent provinces.

The post-1918 period was marked by a concerted effort from the Romanian government to initiate a process of *românizare* (Romanianisation) across all institutions, including academia, to assert national identity and cultural dominance (Stan 2016; Livezeanu 1995: 220–221). Long-standing social and class tensions between Romanian Transylvanians, Jews, and Hungarians were exacerbated by these changes (Pârvulescu and Boatcă 2022: 59). Romanians, predominantly peasants under Austro-Hungarian rule, were regarded as second-class citizens and had limited opportunities to become intellectual elites (Pârvulescu and Boatcă 2022: 53; Hitchins 2002: 81–82). The Romanianisation of universities thus aimed to create Romanian elites to replace the "foreign" ones (Livezeanu 1995: 231–232).

In Greater Romania, there were four universities: in Iași, Bucharest, Cluj, and Chernivtsi. The University of Iași, established in 1860, and the University of Bucharest, founded in 1864, faced significant opposition from students and professors from the outset against "foreigners", especially Jewish students. Jews were excluded from student clubs and had to pay higher tuition fees to be admitted and to study (Livezeanu 1995: 214–218).

The University of Cluj (Kolozsvár), located in the largest city of Transylvania, traces its origins back to the *Academia Claudiopolitana*—the first higher education institution established on the territory of the present-day country—which was authorised to confer academic titles of *baccalaureus*, *magister*, and doctor. In 1872, the Emperor Franz Joseph ratified Laws XIX and XX, formally establishing the Royal Hungarian University in Cluj. On January 4, 1881, Franz Joseph I issued the official founding document of the university and permitted it to bear his name. Before the unification of Transylvania with the Kingdom of Romania, the teaching staff at the University of Cluj was primarily composed of Hungarian-speaking lecturers. After the unification of Transylvania with Romania and the creation of Greater Romania, Decree no. 4090 of September 12, 1919, signed by King Ferdinand I, officially confirmed the "transformation of the Royal Hungarian Franz Joseph University into a Romanian university as of October 1, 1919". The new academic institution comprised four faculties: Law, Medicine, Sciences, and Letters and Philosophy. In its first year, it had over 2,000 students. In October 1927, as a tribute, the Cluj university officially adopted the name of the first king of Greater Romania, thus becoming known as "King Ferdinand I University", a name it retained until 1948.

In the new context, all the Hungarian professors unanimously refused to swear allegiance to the Romanian state, which led to their replacement by Romanian staff (professors) (Karády and Nastasă 2004: 43–44).

An eugenics movement also began to take shape in Romania in the mid-1920s, through the import of ideas from abroad, based on the theory of modern eugenics formulated by Francis Galton in 1883. This theory asserted that all human traits are inherited and that improving the race required efforts of artificial selection. Within the University of Cluj, the principal advocate of Romanian eugenics, Dr. Iuliu Moldovan, a physician and professor of hygiene, developed his own theory of eugenics, which he termed "the hygiene of the nation", along with a theory of biopolitics. These gave rise to a veritable school of thought within the University, representing a blend of pseudo-science and nationalism. The University of Cluj played an essential role in promoting the Romanian element in post-1918 Transylvania. As the only university in the region and with an entirely Romanian teaching staff, it was seen as a realisation of the Union's ideal (Stan 2021: 144), or a Romanian national ideal.

In this context, Sextil Puşcariu, rector of the University of Cluj, emphasised the institution's importance in fulfilling the dream of the "parents and ancestors of ours to have a Romanian university in the heart of Transylvania" (Puşcariu 1921: 1). Although students of various nationalities were accepted at the university, the rising nationalist sentiment led to anti-Hungarian and antisemitic attitudes and actions between the two World Wars. Hungarian students were forbidden from forming their own organisations (Livezeanu 1995: 226), and Jewish students were subjected to violent attacks. According to Pârvulescu and Boatcă, antisemitism in Eastern Europe, and implicitly in Transylvania, was perceived as an issue of Jewish loyalty, influenced by the political and economic dynamics of the region. They argue that the so-called Jewish Question emerged as a reaction to the rise of this urban bourgeois class, perceived as having non-indigenous origins (Pârvulescu and Boatcă 2022: 59; Motta 2019).

The university, driven by nationalist ideology and a history of cultural inferiority complex (Turda et al. 2022), became fertile ground for racist attitudes. In the 1920s, the university witnessed the rooting of the Romanian fascist movement within its walls, with violent student uprisings against Jewish students in 1922, due to a perceived imbalance in the proportion of Jewish students in academia (Livezeanu 1995: 245–46), as well as attacks against Jewish businesses and religious sites (Cârstocea 2014: 46).

3. Sources and Methods

3.1. Historical Sources

The source material underpinning this research consists of the *Anuarele Universității din Cluj* (*The Annals of the University of Cluj*) from the period 1919–1942, which were digitised by the Lucian Blaga Central University Library in Cluj-Napoca. The analysed corpus comprises 18 volumes, published over 23 years, with a total of 6,050 pages. The period 1919–1942 was chosen for analysis because the Romanian University of Cluj was established in 1919 and operated in Cluj until 1942, when it was moved to Sibiu due to territorial changes caused by the war. The annals were published annually and offer a complete record of the official university activities and discourse during the interwar period. They contain speeches by university representatives, information about courses, students, student clubs, committee discussions, and academic activities of each department.

3.2. Data Preparation

The materials were digitised and then preprocessed using Optical Character Recognition (OCR) with ABBYY FineReader, by the Lucian Blaga Central University Library in Cluj-Napoca, to detect text from the scanned images. The PDF files were subsequently converted into plain text files for easier handling. Due to the poor quality of the scans, analysing the university annals proved to be a challenging task. As is common with digitised historical sources, the raw text contains an overwhelming number of OCR errors. Similar to spelling errors, OCR errors are generated through different means and have their own characteristics (Nguyen et al. 2019). First, it was necessary to manually review samples from the source text to identify the most frequent errors, such as extra spaces, new lines, unnecessary symbols, watermarks, and incorrectly divided or concatenated words, which had to be corrected in order to carry out the analysis. After identifying the most common errors, Python scripts were developed to detect and either correct or remove these errors. Using *Regular Expressions* (regex), a series of corrective modifications were applied to the initial texts, which proved to be a reliable method for errors that follow recurring patterns (Volk, Furrer, and Sennrich 2011). Regular expressions are strings of characters that match a pattern within a given text. By using regex, we aimed to find patterns of text errors that could be replaced with the correct text or completely deleted if they were not part of the actual text content to be analysed.

After cleaning, the text was tokenised using Natural Language Processing (NLP) via the *SpaCy* library in Python, meaning it was split into individual tokens or units, such as punctuation marks, words, or groups of letters. Considering only the words, each unit in the corpus was then compared to a custom Romanian dictionary, and efforts were made to merge incorrectly split words and divide incorrectly concatenated words based on custom rules defined in the Python script. Finally, a Romanian spellchecker was used through the *Language Tool* library in Python. The lack of adequate spellchecking tools for Romanian to correct OCR errors made the data preparation process significantly longer and more difficult than anticipated.

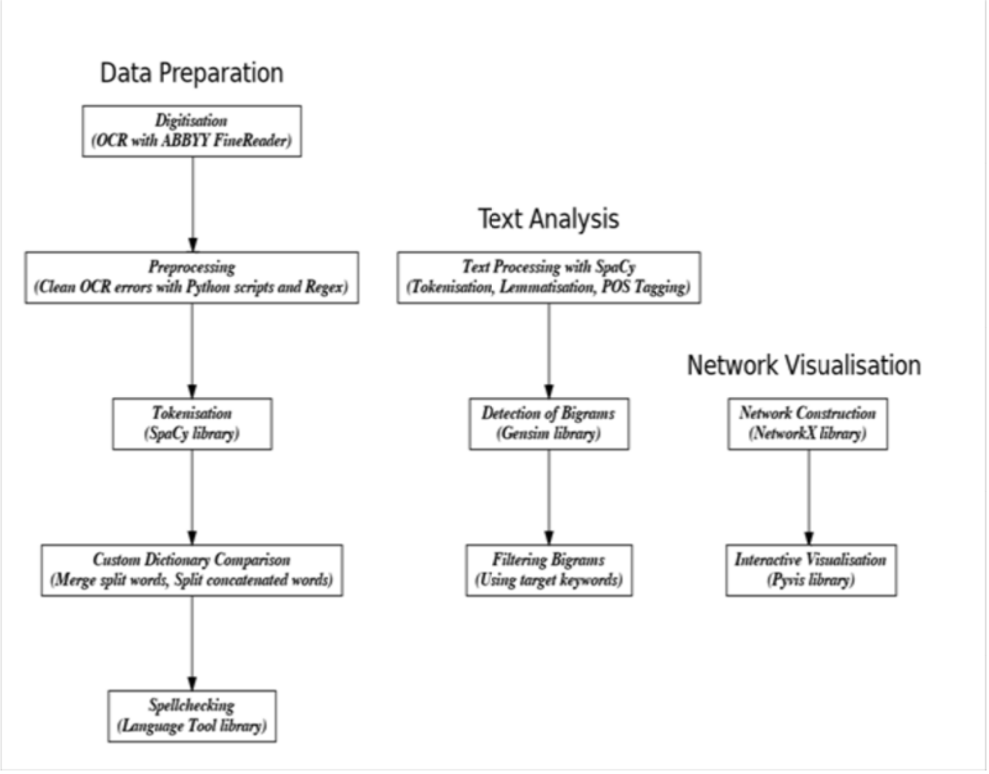
3.3. Text Analysis

3.3.1. Text Processing

"Natural language processing (NLP) is a subfield of artificial intelligence that tries to process and analyze natural language data." (Vasiliev 2020: 20). Natural language refers to language that has been "developed and evolved by humans through natural use and communication" (Sarkar 2016: 2). Natural language can be processed and analysed from several perspectives, namely: phonological, morphological, lexical, syntactic, semantic, discursive, and pragmatic (Khurana et al. 2023).

Therefore, for a computer to be able to analyse texts as data, the text must first be processed. To analyse textual data, in this case we used the *SpaCy* module in Python for advanced natural language processing. "SpaCy is designed specifically for production use and helps you build applications that process and 'understand' large volumes of text. It can be used to build information extraction or natural language understanding systems, or to pre-process text for deep learning." (Honnibal et al. 2020).

Figure 1. Work stages during the analysis phase.


Data Preparation
Digitisation (OCR with ABBYY FineReader)
Preprocessing (Clean OCR errors with Python scripts and Regex)
Tokenisation (SpaCy library)
Custom Dictionary Comparison (Merge split words, Split concatenated words)
Spellchecking (Language Tool library)
Text Analysis
Text Processing with SpaCy (Tokenisation, Lemmatisation, POS Tagging)
Detection of Bigrams (Gensim library)
Filtering Bigrams (Using target keywords)
Network Visualisation
Network Construction (NetworkX library)
Interactive Visualisation (Pyvis library)

First, we split the text into words or tokens, converted each word into its dictionary form, and filtered them according to their part of speech. Thus, the Romanian-language *SpaCy* module (ro_core_news_md) was loaded with only the Tokenisation, Lemmatisation, and Part-Of-Speech (POS) tagging functions activated. The tokenisation function was used to divide the text corpus into individual tokens or units consisting of individual words separated by spaces. In this way, the text was transformed into a list of words that were relevant for analysis, while at the same time removing unwanted characters and filtering the words based on a custom list of excluded terms, punctuation, numerical values, and short tokens that did not serve the purpose of the analysis.

In order to normalise the words used for analysis, the lemmatisation function was applied to reduce all tokenised words to their dictionary form. For example, the lemma of "*românilor*" is "*român*". This allowed us to eliminate duplicate words with slightly different endings and to perform our text analysis on a clean and normalised list of words. The use of part-of-speech tagging enabled us to consider only nouns and adjectives for this analysis, removing words irrelevant to the current study, such as prepositions, conjunctions, verbs, or adverbs.

3.3.2. Bigram Analysis

Next, the *Gensim* library in Python was used to detect bigrams. Bigrams are common phrases or meaningful pairs of words that frequently appear in close proximity. "Gensim is designed to process raw, unstructured digital texts (*plain text*) using unsupervised machine learning algorithms." (Řehůřek and Sojka 2010). For the purpose of this analysis, we applied the *Gensim Phrases* model to the text corpus to detect bigrams or two-word expressions, in order to understand which ideas related to race and national and ethnic identity were most frequently used in proximity within academic discourse between the two world wars. The *Phrases* model automatically detects groups of words that are collocated in a given text string.

To achieve this, a double filtering process was used. First, a custom stopword list was defined to filter out common terms and university-specific terms that were irrelevant to the analysis, such as "institute", "course", "seminar", "director", etc. Secondly, the tokens in the corpus were filtered to include only those containing specific target keywords relevant to our research objective, relating to race, ethnicity, and national identity. The complete list of words used for this bigram identification analysis is as follows: "*rasă*" (race), "*rasial*" (racial), "*rassă*" (the German spelling of the word race), "*superior*" (superior), "*superioritate*" (superiority), "*român*" (Romanian), "*străin*" (foreigner), "*românesc*" (Romanian [adj.]), "*românitate*" (Romanian identity), "*românește*" (in Romanian), "*neam*" (nation/ kin), "*etnic*" (ethnic), "*etnie*" (ethnicity), "*stăpân*" (master), "*evreu*" (Jew), "*evreiesc*" (Jewish), "*evreime*" (Jewry), "*țigan*" (Gypsy), "*țigănesc*" (Gypsy [adj.]), "*țigănie*" (Gypsyness), "*ereditate*" (heritage), "*ereditar*" (hereditary). Thus, only bigrams containing at least one of the words from the above keyword list were considered.

The *Gensim Phrases* model was applied to the cleaned, tokenised, lemmatised, and pre-filtered words from the corpus. Two parameters were used at this stage of the analysis. One parameter was a minimum count of five, meaning that a bigram would be considered only if it appeared at least five times in the corpus. The second parameter used was a minimum score of ten, meaning that only word pairs with a score higher than ten were regarded as bigrams. The purpose of this score is to quantify how often two words appear together as an expression or bigram, to avoid issues related to very rare bigrams and word pairs that may appear by chance. This score is intended to measure how frequently the word pair appears together compared to the probability of each individual word appearing independently. Finally, a list of bigrams with their occurrence counts was exported into a CSV file for further analysis and manual validation, to normalise the bigram list and to check for outliers.

3.4. Network Visualisation

A graphical network visualisation was then created to represent the relationships between the words in the bigrams and the connections between different bigrams. To generate the network, the *NetworkX* library was used, which "is a Python package for the creation, manipulation and study of the structure, dynamics and functions of complex networks" (Hagberg, Schult, and Swart 2008). For interactive exploration of the network, *Pyvis* (Perrone, Unpingco, and Lu 2020) was employed to create an interactive graphical network visualisation, which was exported as an HTML file to facilitate navigation and analysis within a web browser. *Pyvis* networks can be customised both visually and functionally directly from a Python script and can also be operated from the browser, allowing users to change the physical attributes of the graphical visualisation related to the type of engine used for rendering. In this case, the repulsion engine was used, where nodes repel each other to prevent overlap and make the visualisation easier to read.

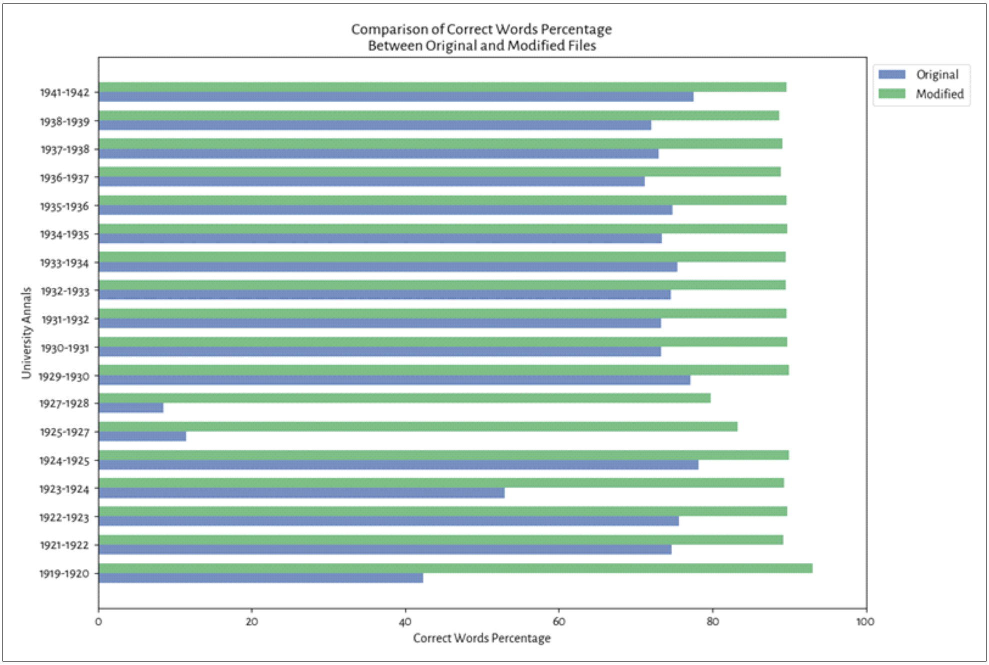
Thus, we constructed a graphical visualisation in which the nodes represent individual words, the size of each node indicating its degree, or the number of connections linked to that node. Nodes that appear in multiple bigrams and therefore have a higher degree will appear larger in the visualisation. The links between nodes represent the bigrams connecting them, meaning that a link is created between two nodes if they form a bigram. The thickness of each link is determined by the frequency count of the bigram.

4. Results

Using *Regular Expressions*, we were able to remove unnecessary marks (such as watermarks), blank spaces, new lines, and metacharacters, while preserving standard punctuation—such as question marks, full stops, commas, etc.—as well as Romanian diacritics. We ensured that a space follows each punctuation mark and adapted the older Romanian punctuation to contemporary linguistic standards (for example, replacing apostrophes with hyphens), so that we could apply modern Romanian spellchecking tools available today.

After completing the OCR error correction process, the quality of the initial texts was improved by over 54%. We calculated this percentage by comparing the words from both the raw OCR results and the corrected texts against a Romanian word dictionary. While the initial texts had a correctness rate of 57%, after the texts were processed and prepared for analysis, we reached a rate of nearly 89% correct words. Although this improvement rate is not absolute, it was sufficiently high to allow us to run the analysis on the corpus and obtain relevant results.

Figure 3. Comparison of the percentage of correct words in the initial, unprocessed texts and in the cleaned and processed texts used for analysis

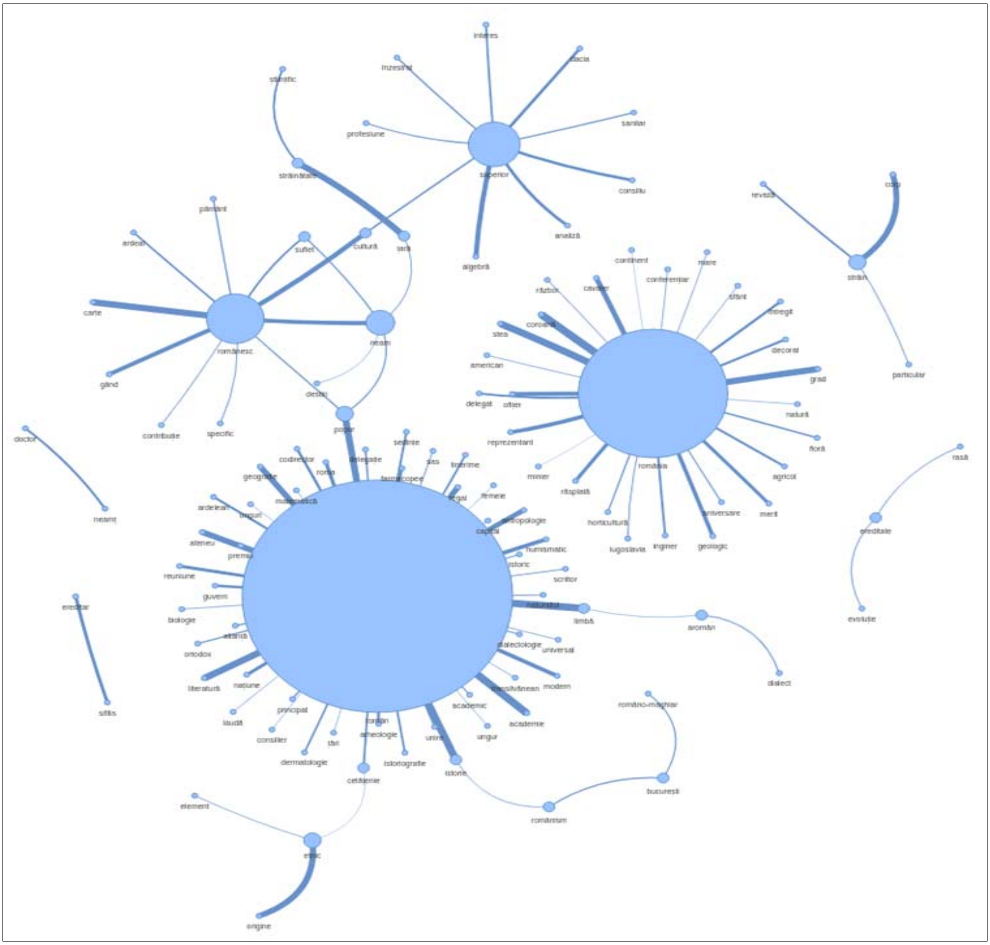


The top ten bigrams by frequency, with occurrences ranging from 62 to 379, are as follows:

Bigram 1 (Romanian)	Bigram 2 (Romanian)	Occurrences	Bigram 1 (English)	Bigram 2 (English)
coroana	România	379	crown	Romania
academie	român	273	academy	Romanian
limbă	român	186	language	Romanian
stea	România	163	star	Romanian
istorie	român	141	history	Romanian
literatură	român	138	literature	Romanian
carte	românesc	130	book	Romanian
România	grad	112	Romania	rank
țară	străinătate	78	country	abroad
popor	român	62	people	Romanian

The graphical visualisation allowed us to more easily observe the connections between the bigrams identified in the corpus. The most significant connections identified with the term "românesc" are: "*carte*" (book), "*cultură*" (culture), "*neam*" (nation), "*gând*" (thought), "*suflet*" (soul), "*pământ*" (land), "*popor*" (people), "*limbă*" (language), "*universitar*" (academic), "*istorie*" (history), and "*modern*" (modern). Two other important connections were identified between the terms "*etnic*" (ethnic) and "*origine*" (origin), and "*străin*" (foreign) and "*corp*" (body).

Figure 2. Text network visualisation



The bigram analysis revealed several key findings. The most recurrent bigram associated with the word "*România*" was "*Coroana României*" (The Crown of Romania), a chivalric order established by King Carol I on 14 March 1881, the day Romania became a kingdom (Clipici 2021: 328). The purpose of this order was to reaffirm the independence and state sovereignty of the new Kingdom of Romania. Another frequent bigram was "*Steaua României*" (The Star of Romania), the oldest and highest Romanian civil order, established on 10 May 1887 as a reward for military and civil services rendered to the state (Clipici 2021: 327). Most professors at the University of Cluj had been decorated with these orders, reflecting their involvement in nationalist activities.

The analysis also highlighted a strong connection between the terms "*românesc*" (Romanian) and "*medical*", indicating the importance placed on the need to reform the medical system in Romania by university intellectuals during this period (Bucur 2002: 188–201). This was also a time when the eugenics movement flourished at the University of Cluj, under the leadership of Iuliu Moldovan, influencing members of departments beyond medicine, such as literature. Key figures such as Gheorghe Bogdan-Duică and Onisifor Ghibu are notable examples of supporters of eugenic ideas and the theory of racial differentiation (Craioveanu 2020).

Another significant finding was the frequent association of the words "*străin*" (foreign) and "*corp*" (body)—*corp străin* (foreign body)—an expression used to describe otherness, appearing 39 times. Similarly, the term "*etnic*" (ethnic) often appeared in proximity to the word "*origine*" (origin), highlighting the polarisation between Romanian identity and other ethnicities. This pairing of terms reflects the strong emphasis placed on Romanian culture, society, and literature, as well as on românitate (Romanianness) in general, within the university.

With few exceptions, Jews were mentioned only in connection with the number of students by ethnicity, and thus the analysis did not return any bigrams containing the word "*evrei*" (Jews). When they were mentioned, it was often in reference to the large number of Jewish students coming from Hungary to study at the University of Cluj, as they were unable to study in Hungary due to a *numerus clausus* policy against Jews. Although a similar *numerus clausus* was proposed at the University of Cluj in an attempt to restrict the number of Jewish students who could enrol, it was never adopted. Other references to Jews pointed to the fact that they were better prepared for the baccalaureate examinations compared to Romanian students, indicating a high level of competition between Jewish and Romanian students. What is noteworthy is that the text frequently distinguishes between Jews and Romanians on the basis of nationality rather than ethnicity, suggesting a conflation or confusion of these two very distinct concepts.

In the early 1920s, violent student attacks occurred at Romanian universities, targeting Jewish students as well as Jewish businesses and places of worship. Among the most fervent supporters of these violent uprisings against Jews in Cluj were students from the Faculty of Law. Regarding the involvement of the Law Students' Society in these student attacks, the university annals note only the following: "As the years 1923-24-25 came, with those turbulent times, this Society too was drawn into the whirlpool of antisemitic movements; its activity waned, its members active throughout the entire student body" (*Anuarul Universității* 1931).

Regarding terms derived from the word "*țigan*" (Gypsy), they are not mentioned at all in the university annals. This omission can be attributed to the history of the Roma population in the Romanian territories, where for nearly half a century they were enslaved, as well as to the status of this population in Greater Romania following emancipation. Many of the most prominent Romanian intellectuals of the interwar period opposed the assimilation of the Roma into Romanian society and regarded them as one of the greatest threats to the Romanian nation (Bucur 2002: 147–48).

The term "*rasă*" (race) frequently appears alongside "*ereditate*" (heredity), particularly in the context of criminology courses taught at the Faculty of Law between 1923 and 1933, which addressed aspects such as heredity in relation to race. Iacob Iacobovici, serving as rector, discussed the "national instinct" in his opening speech for the 1922-1923 academic year, linking it to the concept of race (*Anuarul Universității* 1924: 7). Many courses included lectures on race. For example, an ocular hygiene course taught from 1929 to 1931 discussed the influence of race on ophthalmological issues. Similarly, a course delivered during the 1929-1930 academic year on medicine among primitive civilisations also addressed the concept of race (*Anuarul Universității* 1930: 110).

In the 1929-1930 academic year, a lecture given by the French professor of medical history Jules Guiart was mentioned, in which he discussed the notions of race, stating that "he established that Romanians are almost Gallo-Romans, like our elder brothers, the French!" (*Anuarul Universității* 1930: 112). "In a lecture titled 'Our Sister Romania', Prof. Guiart showed how Romanian blood is imbued with Latinity just like that of the French. Among the Transylvanians, the racial foundation is Celtic, evidenced by their facial features and customs, which are the same as those of Breton peasants. Trajan subdued the Dacians, these Gauls of the East, Caesar the Gauls of the West..." (*Anuarul Universității* 1930: 112-13). Concepts such as the Latin spirit were linked to ideas connected to biology and the human body, such as blood. In the official discourse of the University representatives, a mixture of spiritual and biological terms is observed, with Romania as the common denominator.

Lectures at the Institute of Experimental, Comparative, and Applied Psychology, delivered by C. Gordin, covered topics such as mental evolution, the differentiation of natural types, sex, and race (*Anuarul Universității* 1930: 177). The course on Ethnography and Folklore, taught by Romulus Vuia, included classifications of races and peoples (*Anuarul Universității* 1931: 200;

Anuarul Universității 1935: 273; *Anuarul Universității* 1940: 251-252.). At the Department of Philosophy, Professor Constantin Sudețeanu's courses on Sociology and Ethics encompassed discussions on the physical environment, the biological factors of race, and the realm of religious life (*Anuarul Universității* 1932: 201; *Anuarul Universității* 1938: 317; *Anuarul Universității* 1940: 236).

One of the most prominent figures in both the academic and political spheres, Octavian Goga, spoke about the "feeling of racial differentiation" in his address at the ceremony awarding him an honorary doctorate from the University of Cluj (Goga 1932: 43-44). In social policy courses, Professor Nicolae Ghiulea taught topics such as social-biological policy, hygiene, social medicine, racial defence, eugenics, and the raising of national standards (Ghiulea 1935: 94). Iuliu Moldovan's courses on hygiene and social hygiene included lessons on race within epidemiology, linking it to nationhood, heredity, eugenics, racial hygiene, and segregation (Moldovan 1935: 163-64).

The conference delivered by Gheorghe Popoviciu in 1936, within the Romanian Anthropological Society, focused on the racial similarities and differences between Romanians and neighbouring peoples (Popoviciu 1937: 127). Iordache Făcăoaru published studies on racial diagnosis and composition (Făcăoaru 1937: 157), as well as articles on eugenics and race (*Anuarul Universității* 1939: 176-177; *Anuarul Universității* 1940: 163).

In the academic year 1936–1937, King Carol II was awarded an honorary doctorate and mentioned the concept of race in his speech (*Anuarul* 1938: 91-94). Student papers from constitutional law courses discussed the doctrine of the racist state in National Socialist ideology and National Socialist racism (*Anuarul* 1939: 64). Axente Iancu delivered a course on Mongoloid racial vestiges in Transylvanian children (Iancu 1939: 143), while V. Preda wrote about race and blood groups (Preda 1939: 176-177).

In 1941–1942, Professor V. Papilian lamented the students' lack of interest in race as a political factor, despite its supposed importance in biopolitical sciences (Papilian 1943: 283), while Professor Traian Pop discussed the racial connections between the Dacians and the Romans (Pop 1943: 176-77). Professor Eugen Fischer's 1941 lecture, entitled "Race as a Historical Factor", and Professor Eugenio Morelli's lecture on the racial similarities between Romanian and Italian peasants further illustrate the prominence of racial discourse within the university (Fisher 1943: 189-190).

After analysing the visual representation of the bigrams, it became clear that certain terms and concepts were associated with the notion of Romanian identity, while others were linked to ideas of alterity and ethnicity. The graphical visualisation demonstrates the connections between these terms, providing a starting point for understanding the process of formation and subsequent propagation of strongly nationalist and racist ideologies within the University of Cluj during the interwar period.

5. Conclusions

The history of racism and the eugenics movement in Romania represents a complex and insufficiently explored topic within Romanian academia. This subject reveals a number of important aspects that influenced Romanian society during the interwar period and beyond, with potential impact on later historical epochs as well. As we have seen, racial identity played a significant role during the interwar period, and the eugenics movement, along with sero-anthropological research, had a strong influence on the construction of racial identity in Romania. This led to the formation of a racial hierarchy in which the Romanian people were considered superior to other ethnic groups within the country. This identity construction based on racial criteria had profound implications for collective mentalities and for the policies adopted during the interwar period and the Second World War.

Racism and eugenics played a significant role in the formulation of racial policies during the Second World War, including the deportation of Roma and Jewish populations. This demonstrates that racial ideologies had tangible consequences for vulnerable ethnic groups in Romania. For a deeper understanding of the impact of racial ideologies in Romania, a critical approach is essential—one that explores not only theoretical aspects but also the practical consequences of racism on vulnerable groups and on society as a whole. Understanding the history of racism and the eugenics movement in Romania is crucial for uncovering significant aspects of Romanian history and culture, and for contributing to a deeper and more informed approach to issues related to identity, discrimination, and racial policies.

This study analysed the official discourse within the University of Cluj between 1919 and 1942, as documented in the university annals, employing both automated text analysis methods and traditional textual analysis techniques. Furthermore, the study's findings demonstrate the potential of integrating modern data analysis techniques into historical research. By applying NLP methods to historical text analysis and creating graphical visualisations of the results, we have succeeded in providing a more efficient and accurate approach to examining historical sources. We therefore consider that this interdisciplinary approach opens new avenues for exploring intellectual currents and social networks during the interwar period in Eastern Europe.

The findings of this study reveal significant socio-cultural patterns and highlight the potential of digital humanities to enhance traditional historical research. By integrating modern data analysis techniques, we have sought to provide a deeper insight into the intellectual and social dynamics of the interwar period. The emphasis on Romanian identity and the exclusion of ethnic minorities within the academic discourse reflect broader nationalist ideologies, and understanding these discursive patterns contributes to a more comprehensive view of the complex social and political landscape of the interwar era.

The results of this study highlight the significant role of nationalist and racist discourse in shaping the intellectual landscape of the University of Cluj during the interwar period. The large number of professors decorated with the Order of the Crown and the Order of the Star of Romania indicates that the academic community was deeply involved in nation-building efforts. This strong connection between the academic environment and nationalist endeavours underscores the broader agenda of reforming various sectors, including education and healthcare, to reinforce Romanian ethnic identity.

The frequent mention of race in academic lectures—such as the 1929–1930 course on medicine among primitive civilisations and the conference delivered by the French professor Jules Guiart, who asserted racial similarities between Romanians and the French—further illustrates how race was deeply embedded in the intellectual framework of the time. Concepts like the "Latin spirit" were intertwined with biological terms such as blood, reflecting a blend of spiritual and biological rhetoric, with Romania as the common denominator.

These findings align with the broader historical context of Romanianisation and its impact on the academic environment. The emphasis on Romanian identity and the exclusion of ethnic minorities reveal the university's role in propagating nationalist ideologies. For instance, lectures at the Institute of Experimental, Comparative, and Applied Psychology, as well as courses within the Department of Philosophy, frequently included discussions on race, further embedding these concepts into the academic curriculum. Courses in ethnography and folklore, sociology and ethics, and social policy—which covered topics such as race, eugenics, and the biological factors of social life—illustrate just how deeply entrenched these ideologies were. This integration of race across various academic disciplines influenced not only intellectual discourse but also the education of students and the perspectives of future professionals.

The emphasis on courses and lectures that integrated race and national identity reflects the university's role in promoting the process of Romanianisation and the racist-nationalist discourse among students. This process aimed to replace minority intellectuals with ethnic Romanians and to consolidate a nationalist intellectual environment. The results of the analysis demonstrate how nationalist ideologies became institutionalised within the academic setting, contributing to a broader understanding of the complex interplay between politics, society, and academia in interwar Romania.

The identified connections reflect the prevailing nationalist and antisemitic attitudes at the University of Cluj during the interwar period. These findings align with the broader historical context of Romanianisation and its impact on the academic environment. The emphasis on Romanian identity and the exclusion of ethnic minorities reveal the university's role in spreading nationalist ideologies.

Although the methods used in this study provided valuable insights for the research topic, there are certain limitations associated with this type of analysis. The quality of the scanned source materials complicated the data preparation process for analysis. Additionally, we currently believe it is not possible to obtain a completely error-free text for analysis without manual intervention. Moreover, the lack of adequate OCR tools and spell-checking software for the Romanian language further extends the data preparation phase. However, in future research, these methods can still be applied to other historical sources, such as interwar newspapers and magazines, to examine how the discourse about Romanian identity evolves across different media and connects to ideas such as race, minorities, and national identity.

Acknowledgements

This research was funded by the Doctoral Advanced Fellowships program of STAR-UBB under grant CNFIS-FDI-2023-F-0214, project number 32949/23.06.2023.

We acknowledge the use of OpenAI's ChatGPT in assisting with scripting and code debugging throughout the development of this project.

References

Sources

Anuarul Universității din Cluj. Anul I, 1919-1920. (1921). Cluj.

Anuarul Universității din Cluj. Anul școlar 1922/23. (1924). Cluj.

Anuarul Universității Regele Ferdinand I Cluj pe anul școlar 1929-30. (1930). Cluj.

Anuarul Universității Regele Ferdinand I Cluj pe anul școlar 1930/1931. (1931). Cluj.

Anuarul Universității Regele Ferdinand I Cluj pe anul școlar 1931/32. (1932). Cluj.

Anuarul Universității Regele Ferdinand I Cluj pe anul școlar 1934/35. (1935). Cluj.

Anuarul Universității Regele Ferdinand I Cluj pe anul școlar 1935/36. (1937). Cluj.

Anuarul. 1936-37. (1938). Cluj.

Anuarul Universității Regele Ferdinand I din Cluj. 1937-1938. (1939). Cluj.

Anuarul Universității Regele Ferdinand I din Cluj. 1938-1939. (1940). Cluj.

Anuarul Universității Regele Ferdinand I Cluj-Sibiu în al doilea an de refugiu. 1941-1942. (1943). Cluj.

Secondary sources

Berger, J., Grant, P. (2022). “Using Natural Language Processing to Understand People and Culture.” *American Psychologist* 77 (4): 525–537. https://doi.org/10.1037/amp0000882.

Bucur, M. (2002). *Eugenics and Modernization in Interwar Romania*. Pittsburgh: University of Pittsburgh Press.

Cârstocea, R. (2014). “The Path to the Holocaust. Fascism and Antisemitism in Interwar Romania”. *S:I.M.O.N.–Shoah: Intervention. Methods. Documentation* 1 (1): 43–53.

Cârstocea, R. (2017) “Building a Fascist Romania: Voluntary Work Camps as Mobilisation Strategies of the Legionary Movement in Interwar Romania”. *Fascism* 6(2): 163–195. https://doi.org/10.1163/22116257-00602002.

Clark, R. (2015). *Holy Legionary Youth: Fascist Activism in Interwar Romania*. Ithaca: Cornell University Press.

Clipici, R. M. (2021). “Ordinul Național Steaua României in Grad de Colan. Istorie şi Actualitate”. *Revista Biserica Ortodoxă Română, Buletinul Oficial al Patriarhiei Române* 2: 322–342.

Craioveanu, F. (2020). “Racism in Interwar Romanian Press. Disseminators and Influences in ’Societatea de Mâine’: A Case Study”. *Acta Musei Porolissensis* 42: 151-173.

Dan, P. (2018). “Identity, Collective Memory and Antisemitism”. *Analele Universității Din Bucureşti*, Seria ştiințe Politice 1: 91–106.

Furtună, A-N. (2018.). *E Rroma Rumuniatar thaj o Holocausto: historia, teorie, kultura = Rromii din România și Holocaustul: istorie, teorie, cultură = Rroma from Romania and the Holocaust: history, theory, culture*. Ediție trilingvă. Romane rodimata. Popeşti Leordeni: Dykhta! Publishing House.

Gifu, D., Dascalu, D., Trausan-Matu, S.and Allen, L.K.. (2016). "Time Evolution of Writing Styles in Romanian Language". In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, 1048–54. San Jose, CA, USA: IEEE. https://doi.org/10.1109/ICTAI.2016.0161.

Hagberg, A. A, Schult, D.A and Swart, Pieter, J (2008). "Exploring Network Structure, Dynamics, and Function Using NetworkX". In Varoquaux, G., Vaught, T. and Millman, J. (Eds). *Proceedings of the 7th Python in Science Conference (SciPy2008).* Pasadena, CA USA, pp. 11-15.

Hitchins, K. (2002). "The Idea of Nation among the Romanians of Transylvania, 1700-1849". In *Nation and National Ideology. Past, Present and Prospects.* The Center for the History of the Imaginary and New Europe College, pp. 78–109.

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A.(2020). "spaCy: Industrial-Strength Natural Language Processing in Python". https://doi.org/10.5281/zenodo.1212303.

Ioanid, R. (1990) *The Sword of the Archangel: Fascist Ideology in Romania.* East European Monographs, no. 292. Boulder [Colo.]: New York: East European Monographs ; Distributed by Columbia University Press.

Karády, V., Nastasă-Kovács, L (2004). *The University of Kolozsvár/Cluj and the Students of the Medical Faculty: 1872-1918.* Budapest Cluj: Central European University ; Ethnocultural Diversity Resource Center.

Khurana, D., Koli, Khatter, A.K and Singh, S. (2023) "Natural Language Processing: State of the Art, Current Trends and Challenges". *Multimedia Tools and Applications* 82(3): 3713–44. https://doi.org/10.1007/s11042-022-13428-4.

Koszor-Codrea, C. (2022). "Mismeasuring Diversity: Popularizing Scientific Racism in the Romanian Principalities Around the Mid-Nineteenth Century". *Journal of Romanian Studies* 4(1): 37–56. https://doi.org/10.3828/romanian.2022.4.

Livezeanu, I. (1995). *Cultural Politics in Greater Romania: Regionalism, Nation Building and Ethnic Struggle, 1918-1930.* Ithaca: Cornell University Press.

Lucy, L., Dorottya, D., Bromley, P. and Jurafsky, D. (2020). "Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks". *AERA Open* 6 (3): 233285842094031. https://doi.org/10.1177/2332858420940312.

Matei, P. (2022). *Roma Deportations to Transnistria during WWII: Between Central Decision-Making and Local Initiatives*. AT: Wiener Wiesenthal Institut. https://doi.org/10.23777/sn.0222/art_pmat01.

Motta, G. (2019). "Nationalism and Anti-Semitism in an Independent Romania". *Academic Journal of Interdisciplinary Studies* 8(2): 14–26. https://doi.org/10.2478/ajis-2019-0012.

Neagu, L. M. et al. (2020). "Automated Modeling of Romanian Literary Trends in History Using Topics Over Time and Co-Occurences". Bucureşti. https://doi.org/10.12753/2066-026X-20-019.

Nguyen, T-T-H., Jatowt, A., Coustaty, M., Nguyen, N-Van and Doucet, A. (2019). "Deep Statistical Analysis of OCR Errors for Effective Post-OCR Processing". In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 29–38. Champaign, IL, USA: IEEE. https://doi.org/10.1109/JCDL.2019.00015.

Pârvulescu, A. and Boatcă, M. (2022). *Creolizing the Modern: Transylvania across Empires*. Ithaca ; London: Cornell University Press.

Perrone, G., Unpingco, J. and Lu, H-M (2020). "Network Visualizations with Pyvis and VisJS". *arXiv Preprint arXiv:2006.04951*. https://arxiv.org/abs/2006.04951.

Řehůřek, R., and Sojka, S.(2010). "Software Framework for Topic Modelling with Large Corpora". In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Sarkar, D. (2016). *Text Analytics with Python*. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-2388-8.

Smedley, A., and Smedley, B. D.(2005). "Race as Biology Is Fiction, Racism as a Social Problem Is Real. Anthropological and Historical Perspectives on the Social Construction of Race". *American Psychologist* 60 (1): 16–26.

Stan, A-M. (2016). "Statutul profesional şi public al personalului academic de la universitatea românească din Cluj între 1919-1940". In Nastasă-Matei, I. and Rostás, Z. (Eds). *Alma mater în derivă: aspecte alternative ale vieții universitare interbelice*. Şcoala ardeleană de istorie. Cluj-Napoca Bucureşti: Şcoala Ardeleană Eikon.

Stan, A-M.(2021). "De la separatism regional la centralizare: două proiecte legislative ale universitarilor clujeni privind reforma învățământului superior românesc după 1918'. *Plural* 9(1): 141-157.

Szabó, M. K., Ring, O., Nagy, B., Kiss, L, Koltai, J., Berend, G., Vidács, L., Gulyás, A., and Kmetty, Z. (2020). "Exploring the Dynamic Changes of Key Concepts of the Hungarian Socialist Era with Natural Language Processing Methods'. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 54(1): 1–13. https://doi.org/10.1080/01615440.2020.1823289.

Turda, M. (2008). *Eugenism si antropologie rasiala în România, 1874-1944*. Bucureşti: Fundatia Amfiteatru.

Turda, M. (ed.). (2015). *The History of East-Central European Eugenics, 1900-1945*. London: Bloomsbury.

Turda, M. (2016). *Eugenism şi modernitate. Naţiune, rasă şi biopolitică în Europa: 1870-1950*. Libreka GmbH.

Turda, M., and Balogun, B. (2023). "Colonialism, Eugenics and "Race" in Central and Eastern Europe". *Global Social Challenges Journal* 20: 1–11. https://doi.org/10.1332/TQUQ2535.

Turda, M., Bokor, Z., Pârâianu, R., and Varga, A. (2022). *Războiul sfânt' al rasei: eugenia şi protecţia naţiunii în Ungaria : 1900-1919*. Ediţia a 2-A. Cluj-Napoca: Academia Română. Centrul de Studii Transilvane and Şcoala Ardeleană.

Turda, M., and Furtuna, A. N.(2022). "The Roma and the Question of Ethnic Origin in Romania during the Holocaust". *Critical Romani Studies* 4(2): 8–32. https://doi.org/10.29098/crs.v4i2.143.

Vasiliev, Y. (2020). *Natural Language Processing with Python and spaCy: A Practical Introduction*. San Francisco: No Starch Press.

Volk, M., Lenz, F. and Sennrich, R (2011). "Strategies for Reducing and Correcting OCR Errors". In Sporleder, C., Van Den Bosch, A. and Zervanou, K. (Eds). *Language Technology for Cultural Heritage*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 3-22 https://doi.org/10.1007/978-3-642-20227-8_1.