Linguistics

BRIDGING THE GAP BETWEEN DIGITAL HUMANITIES AND NATURAL LANGUAGE PROCESSING: TEXT CLEANING AND NLP ACCURACY EVALUATION OF A SAMPLE OF 20TH CENTURY ROMANIAN NOVELS[1]

Bogdan VĂTAVU
Lucian Blaga University of Sibiu, Faculty of Letters and Arts, Romania

e-mail: bogdan.vatavu@ulbsibiu.ro

David MORARIU
Lucian Blaga University of Sibiu, Faculty of Letters and Arts, Romania

e-mail: david.morariu@ulbsibiu.ro

Abstract

Although Digital Humanities (DH) and Natural Language Processing (NLP) share common ground, the two academic communities around these disciplines rarely benefit from each other. This study aims to bridge the gap between the two fields of study, by attempting to establish whether off-the-shelf NLP tools are suitable for analysing DH-oriented textual corpora. Our case study focuses on *The Digital Museum of the Romanian Novel* (*DMRN*), a digital archive which includes most of the novels in Romanian literature up until 1947. Currently, the archive is composed of more than 1200 digital files in the PDF format, derived from their original print editions. An optical character recognition (OCR) layer has also been added in the process, which opens the possibility of applying NLP techniques to these texts. On the other hand, OCRed texts are notoriously difficult to process, being laden with spelling errors and a significant amount of noise. Our case study had two main objectives: 1) to devise an automated method for cleaning the large quantities of OCRed literary texts found in the archive DMRN, in order to (2) analyse the texts by means of NLP with two of the currently available off-the-shelf models, namely *spaCy* and *Stanza*. The sample of our study consists of 15 novels published between 1933 and 1947. The extracted texts have been cleaned by employing a custom Python script and further sampled for NLP tasks such as tokenization, part of speech tagging and lemmatization. We then assessed the accuracy of the results of each separate task through manual validation. Overall, our methodology has proven to be efficient and time-saving with regard to the automatic cleaning process, but also satisfactory in what concerns the accuracy percentages of the performed NLP tasks.

Keywords: computational linguistics; part of speech tagging; lemmatization; natural language processing; digital humanities; Romanian language corpus; Romanian literary texts.

[1]Article History: Received: 29.12.2025. Accepted: 12.03.2026. Published: 15.05.2026. No funding was received either for the research presented in the article or for the creation of the article.

1. Introduction

The digital turn in the humanities has produced impressive quantities of text to be mined and distantly read[2]. However, while digitization has indeed provided access to entire corpora of digital texts, the task of analyzing this huge amount of data by computational means has proved to be less straightforward. This is especially true when dealing with digital versions of older print materials, the conversion process being famously error-prone, also adding a lot of noise to the finished digital objects. In order to perform analyses on these texts, they need to be cleaned and to undergo noise removal (the removal of superfluous elements in the text, such as paratextual elements or page numbers). Further issues arise when tools not specifically designed for the digital humanities are employed to this end. Some questions emerge, such as: are these tools suitable for the job? Are they accurate? Moreover, similar research communities, in spite of the many commonalities, often pursue divergent research agendas. This is precisely the case with the Digital Humanities (DH) and Natural Language Processing (NLP) communities, which rarely gain from each other's advancements, despite the obvious overlaps, as McGillivray, Poibeau and Fabo (2020) have observed. In this study we aim to develop a methodology that bridges the gap between the two neighbouring disciplines by testing how off-the-shelf tools designed for NLP fare on a corpus of Romanian literary texts compiled for DH applications.

The purpose of the experiment in this article is, therefore, two-fold: (1) to devise an automated method for cleaning the large quantities of OCRed literary texts found in the archive *The Digital Museum of the Romanian Novel* (Baghiu et al., 2019a; from now on, DMRN) (2) to assess the accuracy of two of the currently available off-the-shelf NLP models. The texts in this digital collection range from the early 19th century to 1947, and are all written in the Romanian language. This fact alone poses notable problems in terms of both cleaning and processing since, during this period of Romanian history, typography (i.e. the technical aspects of printing), language and writing have undergone significant changes. With this in mind, and knowing the limitations of the tools at hand, we concentrated our efforts on the novels printed between 1933 and 1947 (Baghiu et al., 2021; from now on DMRN III). The literary texts in this period are edited in a rather consistent manner, the script employed is Latin[3] and, most importantly, the orthography and language of these documents are more uniform and stable than in previous periods. This makes them better candidates for analysis, as the language employed in them is closer to the language in the corpora used for the training of the language models tested here. Therefore, our aim is first to develop a cleaning methodology that is applicable to the entire DMRN archive, and second to test the accuracy of some of the available off-the-shelf language models (trained on contemporary Romanian language corpora), on a sample extracted from several mid-20th century literary texts.

For the purpose of this experiment, we employed two of the most popular NLP software libraries for the Python programming language: *spaCy*, a commercial open-source software library, specifically its large language model for Romanian[4]; and *Stanza*, an open-source toolkit supporting 66 human languages, that was deployed here with its default model for Romanian, which was trained on the RRT corpus.[5]

[2] Already far from being a methodological novelty in contemporary research, the "distant reading" approach that Franco Moretti (2005) has pioneered in literary studies has to be perceived, as the theorist himself states, as a "reductive" and detail-removing, yet pattern- and cluster-generating method.
[3] Older texts in the collection are written in Cyrillic and transitional scripts.
[4] See Ines Montani et al. (2023). For the Romanian language model, see the spaCy models documentation (Explosion, n.d.-a).
[5] See Qi et al. (2020). For the Romanian language model employed here, see the Stanza model performance (Stanford NLP Group, n.d.-b).

ISSN: 2003-0924

2. The State of the Art in Romanian DH and NLP

Digital Humanities (DH) is a relatively new research method in Romania. In the past decade, numerous calls in the direction of adopting this methodology have been made (Goldiș, 2014; Patraș et al., 2019; Pojoga et al., 2019). Nevertheless, as Ovio Olaru has noticed, "the introduction of Digital Humanities (DH) in Romania seems to focus more on the conceptual background of digital culture and on clichés about the emergence of a post-hermeneutical age in literary theory than on actual DH practices." (Olaru, 2019, p. 30)

If we were to broadly categorize the contributions of the last few years that can be included in the DH paradigm, or that relate to it in one way or another, we would contend that these generally fit into one of these three categories: 1) studies that theoretically and culturally advocate for the adoption of computational analysis of texts with all its challenges[6]; 2) contributions that somehow capitalize on the metatextual dimension of this paradigm, by providing an overview of the digital tools and software that can be used for literary texts in general and for literary texts in Romanian in particular (Ciorogar & Modoc, 2019); 3) studies that exploit these tools by using them in text analyses on digital archives and corpora. Even though, at least in the early years, as it has already been noted, the contributions classifiable in the first two categories outnumbered those qualifying for the third, this gap has lately increasingly narrowed. One of the reasons for this is that in the field of literary studies, there are indeed few studies that employ the methods of DH and computational linguistics to draw broad conclusions regarding a certain corpus of texts. From stylometric analyses on corpora of Romanian novels (Modoc & Gârdan, 2020) to network representations of character interactions in canonical novels (Pojoga et al., 2020), but also sentiment analysis in canonical literary works (Gavrilă et al., 2021) and representations of entirely literary and cultural phenomena, such as the "avant-garde networks of East-Central Europe" (Modoc, 2020), the studies that focus on *DH* tools in action have mainly relied on the discrepancy between the perspectives favouring the literary canon and the approaches aiming for comprehensive analysis of literary corpora.[7] There are also insights that value the implications of DH in the didactic process, such as developing interactive platforms featuring a database of Shakespeare's works, for instance (Nicolaescu & Mihai, 2014).

Finally, the statistical analyses and the quantitative surveys are also worth mentioning. Some thematic and special issues have gathered over the past several years many articles that consider Romanian literature through numbers. Editorial projects that include *The Culture of Translation in Romania* (Sass et al., 2018), which comprises a section specifically addressing quantitative studies[8], *Ruralism and Literature in Romania* (Baghiu et al., 2019b; Borza, 2019; Gârdan, 2019), but also thematic issues such as "Literature between Canon and Archive. New Distant Reading Approaches to the Study of the Novel"[9] or "Distant Reading – o nouă

[6] In addition to the already mentioned studies, we may also add the following: Ursa (2015), Bâlici (2018), Bâlici (2019), Ciorogar et al. (2019), Gârdan and Modoc (2020).

[7] In this context, we should also mention studies such as those by Popescu and Dinu (2008) and Dinu and Sulea (2012). While the first one discusses, from the perspective of computational stylistics, "function words frequencies" as "style markers", the latter addresses "stylome", "authorship attribution" and "pastiche detection" from the perspective of computational linguistics and in relation to the texts and style of Mateiu Caragiale, by using "the frequency rankings of stopwords".

[8] The first section of the volume, "General Analysis and Quantitative Studies" gathers contributions that approach literary and translation studies by using quantitative analysis methods and tools. More specifically, the contributions address the avant-garde in the Romanian cultural space, the translation of novels in Romania during the communist regime and poetry translations in Romanian periodicals between 1990 and 2015 (Baghiu, 2018; Modoc, 2018; Pojoga, 2018).

[9] This thematic issue of the *Metacritic Journal for Comparative Studies and Theory* (2020) has featured articles that addressed, in addition to the studies already mentioned on stylometry and the character network, "the Romanian translations of

ISSN: 2003-0924

paradigmă de cercetare teoretică?" [Distant Reading – A new theoretical research paradigm?][10] have reinforced the DH method in Romanian literary studies and have also proposed comprehensive analyses on large corpora.

However, the lack of tools for Romanian texts or even the lack of clean corpora continues to be recurrent statements in DH studies. Whether one ends up having to do "a very close distant-reading" (Gârdan, 2018) or having to use secondary data, there are still a number of trade-offs that DH studies so far reveal. For instance, although impressive from the point of view of the methods and the tools employed, Neagu et al.'s (2020) research uses secondary data (*The Chronology of the Romanian Literary Life*) – and not the actual literary production – to draw otherwise very interesting conclusions regarding the post-war Romanian literary landscape.

A few articles do focus on corpus creation and text cleaning, but they mainly stem from the area of linguistics and the NLP community. Busuioc and Caragea (2019), for instance, describe their efforts to build a corpus that will be utilized for the second edition of *The Dictionary of the Romanian Language*, which also includes literary texts from the 16th to the 20th century. They also provide details regarding the tools and steps taken in cleaning and tagging their corpus for further analysis. Fewer articles deal with the evaluation of the accuracy of Romanian NLP language models (such as those employed here), or the retraining of these for better results. In this respect, the most thorough evaluation of retrained Romanian language models, on a corpus that also includes literary texts can be found in Vasile Păiș et al.'s study (2021).[11] Moreover, there are some other studies addressing language processing, Romanian language corpora building and annotation projects that are also part of an emerging research field. Mainly, all of them focus on the popularization of some research projects[12], Romanian corpora and resources such as "The Reference Corpus of the Contemporary Romanian Language (CoRoLa)" (Mititelu et al., 2014) – described by the authors as "the largest publicly available corpus of Romanian", comprising more than 1.250.000.000 tokens, a corpus that also contains oral materials and whose written texts are all "automatically sentence-split, tokenized, part-of-speech tagged, lemmatized" –, "Linguistic Linked Open Data (LLOD)" and resources available for the Romanian language, as "The Romanian wordnet" and "The Romanian Reference Treebank" (Barbu Mititelu et al., 2020, pp. 32–33). Among these resources, the last mentioned is the most relevant for our study, since both *spaCy* and *Stanza*, i.e. their models for Romanian, have been trained on "RoRefTrees" – the treebank that is annotated in "Universal Dependencies format".

With regard to the Universal Dependencies (UD)[13] for Romanian and the list of morphological tags, out of all 17 universal POS categories, the "UD Romanian RRT" lists 16 of them, from *ADJ* to *X*[14], and out of the 9524 sentences, 1818 are extracted from literary texts, which is a relevant aspect considering our sample. Finally, since we have tested the accuracies of both *spaCy* and *Stanza* for sentence segmentation, POS tagging and lemmatization, i.e. the "morpho-syntactic" and "lexical linguistic features" (Ion et al., 2018), we should also mention

Shakespeare's Bawdy Puns" and thematic lexical analyses on "work", "travel" and "affective geography" in the Romanian novels.

[10] This collection of texts proposes a mixed theoretical and practical/applied approach to the DH paradigm (Goldiș & Modoc, 2020).

[11] Neagu et al. (2020) have also retrained a language model for their analysis.

[12] For an overview of several recent Romanian language projects of the Romanian Academy, including projects on spoken Romanian, but also on the transliteration from the Cyrillic to the Latin script, see Tufiș and Dan (2018).

[13] "Universal Dependencies (UD) is a framework for consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human languages" (Universal Dependencies, n.d.-c). For the UD used in this article, see the UD for Romanian (Universal Dependencies, n.d.-b).

[14] The entire list of UPOS includes: "ADJ, ADP, AUX, CCONJ, DET, INTJ, NOUN, NUM, PART, PRON, PUNCT, SCONJ, VERB, X", while the number of "lemmas" tagged as "particles", "pronouns", "determiners", "auxiliaries" adds up to 124.

ISSN: 2003-0924

that the RRT treebank was, as specified in the description, manually annotated with respect to POS, while lemmas were automatically annotated.

This being the state of the art in DH in Romania and in the relevant NLP area of research, we consider the present study to be a necessary contribution to the field, not necessarily in itself, but rather through the pathways it opens for further investigation of corpora of OCRed texts by natural language processing.

3. The Corpus: Compilation, Sampling and Methodology

The corpus we used for this analysis is part of the archive that resulted from the *Digital Museum of the Romanian Novel* (DMRN) research project. The aim of the project was to digitize all the Romanian novels published between 1845 and 1947. The justification for this temporal delimitation, but also for the three functional intervals – 1845-1900 (DMRN I), 1901-1932 (DMRN II) and 1933-1947 (DMRN III) – is to be found in some aspects of the historical and political background that directly influenced novel production. The analogue tool used by the DMRN research team in the pre-digitization stage was *Dicționarul Cronologic al Romanului Românesc de la Origini până la 1989* [The chronological dictionary of the Romanian novel from origins to 1989] (Istrate et al., 2004; from now on CDRN), which also lists 1845 as the year of publication of the first Romanian novel. Moreover, the other time frames reflect the quantitative trends[15] of the novel and provide coherence for *distant reading* and statistical analyses. For this reason, an experiment such as ours also carries broader historical and cultural significance. The development of a methodology for automated text cleaning, together with the assessment of the accuracy of two currently available off-the-shelf NLP models, facilitates access to large-scale, exhaustive analyses that can address "the great unread" (Cohen, 2002, p. 23) of Romanian literature. In this context, preparing digital texts for complex automated analyses constitutes an essential preliminary stage for future studies that will examine the DMRN from a diachronic perspective on the Romanian language, as well as from the viewpoints of literary history and cultural studies. Moreover, such an experimental study implicitly advocates for recognizing the value of literary works as components of cultural heritage. This perspective follows the recent analysis by Terian and Baghiu (2025), which examines the link between literary heritage and cultural heritage, using the *DMRN* as a case study.

Quantitatively, the *DMRN* archive comprises more than 1200 digitized novels and represents, therefore, one of the most relevant archives for the digital analysis of literary texts written in Romanian.[16] Regarding the conversion[17] of analogue texts into digital texts, the research team has used *CZUR ET16* and *CZUR Aura* scanners, which were also provided with page tools and OCR engines. The OCRization has been completely inefficient for the 19th-century novels written in transitional script, while for the other novels in the archive there were several factors influencing the accuracy of OCR, that we identified while working on this study and which are listed below. In fact, since they relate both to the scanning process and to the typographic and layout peculiarities, these factors have also influenced us to a considerable extent in defining the number of texts and the time frame we have chosen to analyse and to select the corpus for NLP.

[15] In addition to historical delimitations, there are those related to variations in the novelistic production, such as the "boom" starting in 1933, which was a landmark in the setting up of the archive (see Terian, 2019).
[16] The series of research projects aimed at digitizing the Romanian novel also includes projects such as Roxana Patraș, Ioana Galleron, Lucreția Pascariu, Alexandra Olteanu, Ioana Lionte, Camelia Grădinaru (2020), *HAIRO: Corpus de romane haiducești (1850-1950)*.
[17] The digitization team describes the whole process in detail in Coroian-Goldiș et al. (2019).

Vol. 9 No 2 (2026)
ISSN: 2003-0924

The factors related to scanning *per se* include the non-recognition of the last characters at the end of the line in the case of books with too narrow page margins, which could not be properly scanned or standardized in the pre-processing operations. In addition, there were books whose binding prevented the scanning of linearly positioned pages at a 180-degree angle, the rows being skewed very close to the inner page margins. Last but not least, there are also some novels (or fragments of novels) with less satisfactory OCR, due to the scanners' light not being adapted to the type of paper (to glossy paper, for example). Beyond the scanning process, the OCR was also influenced by some factors related to typesetting and page design. The images on the page in the children's novels influence the quality of the OCR as they affect the text layout. The drop caps also reduce its recognition and the accuracy of OCR in nearby characters. The spacing between paragraphs and, even more, the noise are mostly misrecognized and incorrectly linked to certain characters or punctuation. Library stamps or other notations on the pages (which include the page header designs being incorrectly recognized as text lines) contribute with more noise and are embedded in the text of the novels. In addition, many of the ellipses have a negative effect on OCR, especially for characters on the line they occur on. Generally, low quality OCR, even in small percentages at the level of entire documents, is found culpable for significant error rates on downstream NLP tasks, as Mieskes and Schmunk (2019) have observed.

Mainly by noting that the level of OCR accuracy is not satisfactory with novels written in the 19th century, due to the transitional script, but also due to the overall aspect of the scanned pages and the noise that affected the OCR, we have decided – as we have already mentioned – to focus on the last part of the DMRN archive, i.e., the DRMN III, which includes novels published between 1933 and 1947. We have also opted for this time frame because the scope of this study is to pilot a series of studies aiming to optimize the automated cleaning process and the NLP analysis for the 19th century and early 20th century novels that are part of the DMRN archive. Quantitatively, taking into account that these 15 years amass – at the scale of the archive – more than 700 digitalized novels (60% of the total), we have decided to sample this segment of the archive. However, we consider that this sample is representative for the entire DMRN III digital collection.

More precisely, out of this archive of 715 novels[18] we have sampled a number of 15 texts[19], which have become the dataset of our experiment. The method employed here was *stratified random sampling*.[20] The rationale at the basis of establishing the strata was the coverage of the entire 15-year period of DMRN III. Therefore, we have written a Python script which made the random selection of one text per year. A different, and possibly better approach in establishing the strata for the sampling would have been to classify the texts according to the quality of the OCR (a limit which we acknowledge below), but this was beyond the possibilities of such a small research team. Nevertheless, our method resulted in what we consider to be a balanced corpus, comprised of both low-quality prints (which usually result in low quality OCR texts), and some better-quality ones. Also, from the standpoint of the literary subgenres found in the archive and according to the subgenre labels listed in *CDRN*, our texts

[18] This total number of novels represents, according to *CDRN*, approximately 80% of the novel production between 1933 and 1947.

[19] The texts sampled for our analysis are, in chronological order: Jean Bart, *Europolis* (1933); Damian Stănoiu, *Parada norocului* [The parade of fortune] (1934); Take, *Nunta Lizetii* [Lizeta's wedding] (1935); Ariana Del-Vet, *Două lumi* [Two worlds] (1936); Ioan Missir, *Fata moartă* [The dead girl] (1937); G. M. Amza, & Al. Bilciurescu, *Vampirul* [The vampire] (1938); Eugeniu Est, *Zaza* (1939); Liviu Rebreanu, *Amândoi* [Both of them] (1940); Ilie Anesiea, *Zmeu de mare* [Sea dragon], vol. I (1941); D. V. Barnoschi, *Stăvilare. Se aleg apele* [Floodgates. The waters clear] (1942); Cezar Petrescu, *1907*, vol. III, *Pământ, mormânt…* [Land, grave…] (1943); Mircea Streinul, *Băieți de fată* [Illegitimate sons] (1944); Mihail Florescu, *Voluntarii* [The volunteers], vol. I, *Spre Spania* [Toward Spain] (1945); Octav Dessila, *Porți fără număr* [Unnumbered gates], vol. II (1946); Sandra Cotovu, *Vijelie* [Storm] (1947).

[20] For the various sampling techniques employed in NLP, see Xiao (2010, pp. 151–153).

ISSN: 2003-0924

cover a wide range: the "social novel" (2), "sentimental novel" (2) and the "sensation novel" (2), while the other nine texts are classified in *CDRN* as "family", "war", "noir", "detective", "children and young adult", "uprising", "political" "slum" and "erotic" novels. Based on the overview table set up by Andrei Terian et al. (2021), with regard to the diversity of novelistic subgenres in the DMRN archive, the texts we have sampled comprise about 50% of all the subgenres recorded in the table. Moreover, some of them, i.e. the "family", "political" and "slum" novels, comprise subgenres that do not amount to more than 2% and which, according to the same quantitative analysis, have undergone a significant decline compared to previous decades.

An editorial overview on the randomly selected texts represents another important aspect to be discussed, since our study is also reliant on the quality of the prints, as we have already stated. Considering that the mapping of novelistic "production centres" (Morariu et al., 2021) has already revealed the supremacy of Bucharest as the publishing centre where 92% (Morariu et al., 2021, p. 36) of the novels in the DMRN archive released between 1933 and 1947 were published, it was not surprising to find that – apart from *Băieți de fată* [Illegitimate sons] (1944) by Mircea Streinul – which was printed in Craiova, at the ABC Publishing House, all the other 14 texts were published by Bucharest publishing houses. However, the fact that Craiova is listed as the city of publication for one of the selected novels is particularly important, as Craiova only accounts for less than 10% of the Romanian editorial output of the period (while Bucharest accounts for roughly 90%). Yet, against this obvious geographical clustering around the capital city, only Cartea Românească publishing house occurs three times in the list of the 15 selected novels, alongside publishers such as Adevărul, Universala Alcalay, Scânteia and so on.

Two more remarks are needed concerning the relevance of our dataset sample. The first one refers to the fact that all the scanned versions we have worked with from the DMRN archive are digital reproductions of the first editions. This is important for the purpose of our analysis, which is strongly centered on developing a cleaning methodology and testing the accuracy of the *spaCy* and *Stanza* models on the novels published during the 15 years we have selected and on the Romanian language of that time. The second remark concerns the authors of our texts and their level of popularity in the book market at the time, but also their general status in Romanian literature on a broader temporal axis. More precisely, the list of authors in our sample is quite diverse, when placed on an axis from "hypercanonical" (Damrosch, 2006) and "prolific" to non-canonical writers[21], including authors such as Liviu Rebreanu, who belongs to the "hypercanonical group" of Romanian authors, but also Cezar Petrescu, Damian Stănoiu and Mircea Streinul – all of them competing for the leading position in the ranking of the most prolific authors – to Jean Bart and Octav Dessila – who emerge on the literary map of the time as "pioneers", from both a thematic and a narratological perspective. Finally, it includes such authors as Take, Ariana Del-Vet, Ioan Missir and Ilie Anesiea, who have not only authored one single novel each in the entire DMRN archive (1845-1947), but who have always been considered as non-canonical. Bearing these considerations in mind – although our research does not advance hypotheses, nor objectives related to computational stylistics, or to the way the peculiarities of an author's literary style might impact the accuracy of OCR (especially in the realm of punctuation) – we argue that our sample is definitely representative at the scale of the entire digital collection.

[21] For a comprehensive discussion on categorisation as well as concepts, methodology and quantitative data, see Morariu et al. (2021, pp. 36–37).

ISSN: 2003-0924

3.1. Text Extraction and Cleaning

In order to perform NLP tasks on our corpus, we needed to extract the OCRed text from the PDF files. The tool employed for this task was the command-line utility *pdftotext*, part of the *poppler-utils* package in Fedora Linux (Poppler, n.d.). The output was a batch of 15 text files, in UTF-8 encoding.

Our next task was making these text files suitable for natural language processing. With the greater objective in mind – specifically, performing NLP on the entire DMRN archive – we developed a methodology for a largely automatic cleaning process[22] that we can ultimately apply to greater quantities of PDF files. Some cleaning tasks could not be automated and were consequently performed manually. In any case, these do not involve a significant workload and are in accordance with our general objective of cleaning the entire DMRN archive in a timely and effective manner. In the next paragraphs, we detail the entire procedure.

The first step in this process is commonly called *text sectioning* by NLP specialists, and it involves removing the undesirable parts of a digital text (Palmer, 2010, p. 10). In practice, this means discarding such elements as page breaks, page numbers, soft hyphens[23], various page decorations and section markers (such as "***"), along with elements of paratext such as the name of the author or the title of the novel in the header section, prefaces, title pages etc. For this purpose, we have written a *Python* script which employs *Regular Expressions* (RegEx) to automatically remove most of the above enumerated items. The script was highly successful in eliminating page breaks (U+000C); page numbers; soft hyphens (U+00AD) and any remaining hard hyphens; rows of text which only contained newline characters (U+000A), or rows formed only of a single character (generally noise produced during OCR); page decorations that were rendered as sequences of uninterrupted graphic characters during OCR; Roman numerals used for chapter numbering, section breaks and the spaces before punctuation marks, which seem to have been used by certain publishers at the time.[24]

The same script was only moderately successful in removing the name of the author and the title of the novel from the header section of the novels where such elements were present. Many names and titles in our sample dataset (and in the DMRN archive) contain Romanian special letters, such as *ș, ț, ă, î, â*. These letters are very often rendered by OCR engines as graphically similar Unicode characters. In our case, they were usually rendered as their Basic Latin Unicode block equivalents. This situation makes find and replace operations significantly difficult. Therefore, in order to increase the accuracy of the method, we have applied normalization[25] to the matched strings and encoded all of them in the ASCII character set, so that all special characters in Romanian were rendered as basic Latin characters. As a result, the *RegEx* employed was more adept at matching the names of authors and novels from the stem of each individual text in our sample. Nevertheless, the aesthetic decisions of the publishers (e.g.: spacing each individual letter in the header section) and subpar OCR, which frequently misidentifies characters in person names, has hampered the performance of the script in this regard.

[22] Cleaning a corpus of texts can mean very different things, from producing exact copies of the original document, to "the omission of data irrelevant to the intended use of the corpus". In this article we employ this last meaning, as listed by Prinsloo et al. (2022, p. 166).

[23] Hard hyphens, which are employed (like soft hyphens) to separate words on two different lines of text, were dealt with by *pdftotext*, in the process of text extraction. The tool also unites the lines of text that were separated in this manner.

[24] However, there are also some aspects that affected the post-OCR operations and the cleaning process that we noticed once we ran the script. The removal of page numbers and the removal of paratextual elements were not automatically done when the last line on the page ended with a split word, since replacing the hyphen caused the syllables of that word to be merged with the page number, author's name or title. Moreover, having the titles in the page header written in a different font or with larger spaces between characters has also prevented their removal in the automatic cleaning process.

[25] For a defition of normalization, see the *Unicode normalization FAQ* (Unicode, n.d.).

ISSN: 2003-0924

As for the actual text of the novels, we restricted our interventions to the minimum, keeping intact the spelling employed at the time. The only exception made concerns the apostrophe, which was used, based on the linguistic rules in effect before the 1953 spelling reform (Vintilă-Rădulescu et al., 2021, p. 56), to indicate the dropping of certain sounds or syllables and their omission. It is also well known that in contemporary Romanian the hyphen has largely taken over the role of the apostrophe (Vintilă-Rădulescu et al., 2021, p. 56). Replacing all apostrophes in our texts with hyphens seemed to be an intervention worth making, as it improved accuracy in POS tagging in our preliminary tests. Also, this change does not affect the lexical and syntactical integrity of the texts in our corpus.[26]

With regard to the manual cleaning tasks that we have already mentioned, they may be classified based on the frequency with which they have been performed in texts as follows. First of all, we have manually removed the titles from the header section which, due to the printers' aesthetic choices, were overlooked by the *RegEx*. This intervention has been required in 10 out of the total of 15 texts from our sample, but we should also mention that the titles occurred irregularly throughout the text, leading us to conclude that the *RegEx* employed has removed at least some of these paratextual elements. Two relevant examples for this trend of misidentifying the titles of texts are provided by the novels *Vampirul* [The vampire] (1938) by G. M. Amza and Al. Bilciurescu and *Zaza* (1939) by Eugeniu Est, whose titles are both written in spaced capital letters: "V A M P I R U L" and "Z A Z A". The same intervention was carried out for authors whose names were also manually removed if they had not been matched by our script and for the Roman numerals indicating the chapters (and even for the word "chapter" if it was used with the numbers). Finally, our manual cleaning also included some minor interventions whenever the page numbers were merged with words separated at the end of the line (as a result of the automated removal of hard hyphens). Even though we have not performed any intervention on the texts themselves, we have noticed some recurrent errors that we also intend to remove through automatic cleaning, so that we can reproduce the entire process for a larger amount of documents, as already mentioned. One of these errors occurs in most situations where the hard hyphen precedes the third person singular masculine form of the pronominal clitic ("-l"), which is misidentified as numeral *1* ("-1").

After this step, we employed a regular expression in a Python script to eliminate all the newline characters in the sample texts, which seemed to improve the accuracy of the NLP tasks in our preliminary tests. The end result was a batch of 15 text files, each one of them a single line of uninterrupted Unicode characters.

3.2. NLP Accuracy Evaluation

Next on our list was assessing the accuracy of some off-the-shelf NLP Romanian language models. Since this operation is manual and time consuming – and taking into account the fact that the research team is formed only of the authors of this article themselves – we proceeded to sample our text files to produce a manageable dataset. In order to obtain a representative sample, we again employed *stratified sampling*. The text files were consequently split into three sections, beginning, middle and end, based on the number of sentences in each of them. This step therefore coincided with the first NLP task we had in mind, which was sentence segmentation. From each of the aforementioned sections we randomly extracted 10 sentences, which amounted to 30 sentences/text. This process was performed with both *spaCy*

[26] With the exception of the many phrases in French, which were somewhat abundant in the literary texts of the period. Being that the NLP tasks we wanted to perform are specifically targeted at the Romanian language, we considered that the alterations produced in this instance do not affect the scientific value of our experiment.

ISSN: 2003-0924

and *Stanza*, resulting in two CSVfiles,[27] each containing 450 sentences, which were then manually assessed for accuracy.

The total number of tokens in both CSV files amounted to 14.845 (7.130 in the *spaCy* corpus and 7.715 in *Stanza*'s). Manually evaluating the accuracy of each one of them for the tasks we had in mind – namely OCR, part-of-speech tagging (POS) and lemmatization – was therefore beyond the possibilities of our small team. Consequently, a similar process of sampling as the one previously mentioned was carried out (extracting 15 sentences/text), resulting in a total of 7.599 tokens, split between 30 individual CSV files, two for each of the selected novels (with outputs from *spaCy* and *Stanza*, respectively). We, therefore, established our two datasets, on which we performed the already stated NLP tasks.

With regard to sentence segmentation, both models have yielded similar results, as can be seen in *Table 1*. *spaCy* scores an average accuracy of 87,38%, the lowest percentage being 66,67 for the novel *Vijelie* (1947) by Sandra Cotovu – the most recently published novel in our sample –, while two novels, *Două lumi* [Two worlds] (1936) by Ariana Del-Vet and *Porți fără număr* [Unnumbered gates], vol. II (1946) by Octav Dessila scored the maximum percentage. The entire series of errors we have noticed when manually validating the sentence segmentation could be discussed as OCR and punctuation errors as well as errors relating to the transition from direct to indirect speech. In fact, these types of errors are connected, since the low accuracy of the OCR affects the recognition of punctuation. More precisely, there are many examples of noise leading to mis-segmentation, as in those cases where the OCR detected periods within the sentence and, as a result, the segmentation was done right in the middle of the sentence. There were also cases of other punctuation marks – such as commas – or of other diacritical marks being misidentified as periods and, consequently, leading to the same errors. Moreover, *spaCy*, as well as *Stanza*, has also segmented sentences consisting only of punctuation marks such as ellipses or brackets.

Title	Author	Publication year	spaCy	Stanza
Europolis	Jean Bart	1933	76,67%	93,33%
Parada noroculuі	Damian Stănoiu	1934	83,33%	90,00%
Nunta Lizetii	Take	1935	96,67%	96,67%
Două lumi	Ariana Del-Vet	1936	100,00%	93,33%
Fata moartă	Ioan Missir	1937	86,67%	90,00%
Vampirul	G. M. Amza, Al. Bilciurescu	1938	90,00%	86,67%
Zaza	Eugeniu Est	1939	86,67%	93,33%
Amândoi	Liviu Rebreanu	1940	90,00%	83,33%
Zmeu de mare, vol. I	Ilie Anesiea	1941	73,33%	73,33%
1907, vol. III, Pământ, mormânt…	Cezar Petrescu	1943	93,33%	83,33%
Stăvilare. Se aleg apele	D. V. Barnoschi	1943	76,67%	93,33%

[27] A Comma-Separated Values (CSV) file is a simple text file format used to store tabular data, where each value in a row is separated by a comma.

ISSN: 2003-0924

Băieți de fată	Mircea Streinul	1944	93,33%	83,33%
Voluntarii, vol. I Spre Spania	Mihail Florescu	1945	86,67%	90,00%
Porți fără număr, vol. II	Octav Dessila	1946	100,00%	90,00%
Vijelie	Sandra Cotovu	1947	66,67%	80,00%
		Average total	**87,38%**	**87,62%**

Table 1. *Sentence segmentation*

Also, with respect to punctuation,[28] two other aspects that have led to errors in sentence segmentation using *spaCy* should be pointed out. Firstly, the case of ellipses deserves a separate discussion, considering the frequency of their occurrence in literary texts. The issue that affects OCR is that, due to their role (i.e., to mark an "interruption" of discourse), ellipses can also occur within the same sentence[29] as in: "my dad hit me… with his whip…accidentally…"[30]. This is a very confusing scenario for automatic segmentation. Similarly, the colon also requires further clarification, since one of its prevalent roles in literary texts involves the transition from direct to indirect speech (Graur, 2009, p. 38). In other words, it is quite obvious that the automatic segmentation reveals this ambiguity caused by colons introducing additional information within the same sentence and all the cases where the same punctuation mark was used to indicate the transition from indirect to direct speech. On the other hand, *Stanza* scores an average accuracy of 87,62%, which is very similar to the average percentage provided by *spaCy*. In addition to the error patterns, we have already highlighted in the case of *spaCy*, *Stanza* has also attested that sentence segmentation is sensitive to the quality of the OCR and the resulting noise. More specifically, a large amount of our invalidated results features instances of misidentification and confusion between punctuation marks (even the end-of-line hyphen being misidentified as period).

As for the accuracy of OCR, POS tagging and lemmatization, it should be noted that we have invalidated both POS tagging and lemmatization every time we have invalidated the OCR (i.e. we dismissed the correct matches as serendipitous). However, all the other cases of errors or ambiguity have highlighted some aspects regarding the limitations of these morphological and lexical taggings. Before listing some representative categories for these errors we have encountered, it may be worth adding that the "Statistics of UD Romanian RRT", i.e., the Universal Dependencies Romanian Reference Treebank, section comprises further considerations related to "the 10 most frequent ambiguous lemmas" and "the 10 most frequent ambiguous types", and also a list of examples, for all the POS tags.[31] The first category (which numbers, nevertheless, only a few errors) can be attributed to the morphological and lexical incompatibility between the contemporary Romanian treebank (RRT) and the vocabulary used in the novels from our sample which includes archaic, vernacular or non-standard variants of some words. For instance, the use of the popular future tense in Romanian – „am să viu" (I will

[28] The difficulties that punctuation poses in sentence segmentation make the topic of David Palmer's study on "Tokenisation and Sentence Segmentation", which addresses, among other issues and challenges in the development of algorithms, the four types of "dependencies" that should be considered – "language dependence", "character-set dependence", "application dependence" and "corpus dependence" –, but also the way a language's punctuation system influences the delineation of sentence boundaries. Moreover, the problematic aspect of punctuation in Romanian is that, as Nunberg (1990, p. 9) stresses with respect to English, there is a lack of studies addressing this issue, i.e. the "punctuation system" in use, not only from a normative point of view, but also from a "descriptive" one.

[29] As outlined in Hanu et al.'s (2018, p. 99) statistical and corpus-based study on Romanian orthographic and punctuation systems, Alexandru Graur (2009, p. 39) also addresses the relative placement of ellipses.

[30] "M-a lovit tata… din greșeală… cu cravașa". Our translation from *Zmeu de mare*, vol. I (1941) by Ilie Anesiea, 13.

[31] See the ambiguous lemmas and the ambiguous types for each POS tag separately in *Statistics of UD Romanian RRT* (Universal Dependencies, n.d.-a).

ISSN: 2003-0924

come) –, „viu" being a popular version for the verb „a veni" (to come), leads to *Stanza*'s tagging of this token as *ADJ*, considering the homonymy with the masculine singular form of the adjective „viu" (alive). Related to these confusions, we should also mention the errors we encountered when validating the *AUX* and *VERB* tags in *spaCy*, since it listed as *AUX* many other verbs besides the three lemmas – "avea, fi, vrea" – which were tagged as *AUX,* when they were not tagged as *VERB*. Moreover, there are also some confusions of POS tags caused by morphological homonymy, concerning the coordinating conjunction „iar" (an adversative conjunction tagged only as *ADV*, since it figures only in the list of "10 most frequent *ADV* lemmas and types" in the treebank statistics), or some *PROPN* tagged as *NOUN*s or *ADJ*s, as "Lina" or "Roza", which also occur as adjective (the former) or common noun (the latter) in Romanian. We should also mention some similar cases of verbal forms, such as "*trebue" ("trebuie" – must), "*înnăbușe" ("a înăbuși" – to smother), "*înnoate" ("a înota" – to swim), "*mângâe" ("a mângâia" – to caress), or adverbial forms, as "*alaltăer" ("alaltăieri" – the day before yesterday), to mention just a few, which yield errors of lemmatization due to misspelling[32] or incorrect grammar. Last but not least, grammatical conversion also leads to a series of errors in POS tagging, to which we should add several examples of grammatical gender and number confusions – "avere" (fortune) with the lemma "*averă"; "petale" (petals) – "*petal"; "cucoane" (ladies) – "*cucoan" – and other inconsistencies between some articulated and unarticulated nouns or different parts of speech – "cafenea" (café) – "*cafene"; "frumoasele" (the beauties) – "*frumoas", "cafeaua" (the coffee) – "*cafeau", "dânsul" (*he* as Romanian polite pronoun) – "*dâns".

4. Results

The results of this manual evaluation enabled us to draw some initial conclusions regarding the performance of the *spaCy* and *Stanza* models for Romanian and see whether there are any correlations between OCR accuracy and the success rate of POS tagging and lemmatization. We will first outline these conclusions separately for each model, followed by some conclusions drawn from a comparative analysis of the results.

The method employed for the evaluation of the different NLP tasks, as well as for the performance of the OCR engine, is accuracy: manually validating *true positives* and *false positives* in our corpus. We agree that employing *F-Score* would have been a much more reliable measurement method for the accuracy of the OCR.[33] Nevertheless, this method necessitates what NLP practitioners call a *gold standard*, a manually annotated corpus against which one can evaluate accuracy. Nevertheless, in DH, as McGillivray, Poibeau and Fabo (2020, para. 37) have observed, "there are no gold standards in the same sense intended in NLP". Creating such a gold standard, whether possible or not, was in any case beyond the scope and possibilities of this small research team. For this reason, we settled for the less-than-optimal method of measuring basic accuracy, a considerable limitation of our study, to which we will return in the following sections. OCR accuracy here should therefore only be taken into account with regard to our corpus, and not for the entire DMRN III archive.

As detailed in the previous section, in terms of segmentation, any of the two models seem to be equally accurate on our samples. Surprisingly for us, sentence boundary detection was actually more challenging for the tested language models than the more specialized tasks

[32] There are also instances of correct recognition of lemmas despite misspellings. See the example of the noun "femeia" (the woman), misspelled "femeiea", for which the lemma provided was "femeiee", with the correct unarticulated forms being identified.

[33] For example, Păiș et al. (2021, pp. 337–338) have used *F-Score* to measure the performance of *tokenization* and *sentence boundary* detection, and basic accuracy for POS and lemmatization.

ISSN: 2003-0924

of POS tagging and lemmatization. Given the literary character of our corpus, we expected the former to be more difficult and error-prone for the off-the-shelf models employed here. But our experiment actually shows that the apparently simpler task of sentence segmentation poses specific difficulties when it comes to literary texts. This result is also in accordance with other studies, which remark on the general low performance of sentence segmentation on otherwise good quality OCRed text, if punctuation is misrecognized or supplemented by OCR noise (Van Strien et al., 2020, p. 488).

On the other hand, POS tagging and lemmatization have proved to be surprisingly good, especially with *Stanza*. As shown in *Tables 2* and *3*, *Stanza* outperforms *spaCy* at POS tagging for all the texts in our corpus (if one counts the greyed-out areas in the table). *spaCy* does seem to occasionally perform better than *Stanza* at lemmatizing (in 4 distinct cases), but overall *Stanza* does the better job at this task, as well, in 11 out of our total of 15 samples. *Stanza* also barely yields percentages lower than 90 on our corpus of texts, which makes it the better candidate of the two for future DH research on the DMRN III archive.

Title	Author	Publication year	No. of tokens	OCR	POS	Lemma
Europolis	Jean Bart	1933	156	98,72%	88,46%	94,23%
Parada norocului	Damian Stănoiu	1934	220	97,27%	86,64%	88,48%
Nunta Lizetii	Take	1935	263	99,62%	91,57%	94,64%
Două lumi	Ariana Del-Vet	1936	241	99,59%	86,72%	96,27%
Fata moartă	Ioan Missir	1937	247	93,93%	87,45%	89,88%
Vampirul	G. M. Amza Al. Bilciurescu	1938	256	97,66%	90,23%	92,58%
Zaza	Eugeniu Est	1939	77	98,70%	87,01%	92,21%
Amândoi	Liviu Rebreanu	1940	234	99,57%	91,45%	92,74%
Zmeu de mare, vol. I	Ilie Anesiea	1941	137	90,51%	80,15%	84,56%
Stăvilare. Se aleg apele	D. V. Barnoschi	1942	186	98,39%	95,70%	93,55%
1907, vol. III, Pământ, mormânt…	Cezar Petrescu	1943	132	99,24%	81,82%	85,61%
Băieți de fată	Mircea Streinul	1944	166	96,39%	80,12%	87,95%

Vol. 9 No 2 (2026)
ISSN: 2003-0924

Voluntarii, vol. I Spre Spania	Mihail Florescu	1945	186	99,46%	93,01%	94,09%
Porți fără număr, vol. II	Octav Dessila	1946	175	82,86%	67,24%	71,84%
Vijelie	Sandra Cotovu	1947	187	94,65%	85,03%	86,63%
Tokens total and accuracy averages:			**2863**	**96,61%**	**86,80%**	**90,16%**

Table 2. *Accuracies* spaCy *(punctuation not included). Grey cells show where this model outperformed the other.*

Title	Author	Publication year	No. of tokens	OCR	POS	Lemma
Europolis	Jean Bart	1933	188	96,28%	93,62%	89,36%
Parada noroculuí	Damian Stănoiu	1934	243	99,18%	95,88%	94,65%
Nunta Lizetii	Take	1935	201	98,51%	96,52%	95,52%
Două lumi	Ariana Del-Vet	1936	186	97,31%	94,62%	92,47%
Fata moartă	Ioan Missir	1937	208	92,31%	91,35%	89,42%
Vampirul	G. M. Amza Al. Bilciurescu	1938	268	97,76%	95,15%	94,03%
Zaza	Eugeniu Est	1939	111	99,10%	90,99%	93,69%
Amândoi	Liviu Rebreanu	1940	237	96,20%	94,51%	94,09%
Zmeu de mare, vol. I	Ilie Anesiea	1941	153	94,12%	90,20%	90,85%
Stăvilare. Se aleg apele	D. V. Barnoschi	1942	304	97,37%	96,05%	93,42%
1907, vol. III, Pământ, mormânt…	Cezar Petrescu	1943	150	96,00%	90,67%	89,33%
Băieți de fată	Mircea Streinul	1944	214	98,60%	93,93%	92,52%
Voluntarii, vol. I Spre Spania	Mihail Florescu	1945	203	98,52%	96,06%	95,07%
Porți fără număr, vol. II	Octav Dessila	1946	290	93,79%	91,03%	91,38%

ISSN: 2003-0924

Vijelie	Sandra Cotovu	1947	254	95,67%	92,91%	92,52%
Tokens total and accuracy averages:			**3210**	**96,67%**	**93,80%**	**92,68%**

Table 3. *Accuracies* Stanza *(punctuation not included). Grey cells show where this model outperformed the other.*

As for any correlations between poor OCR and low NLP performance, things are not that straightforward. *spaCy,* again, fares worse in this respect than *Stanza.* Some of our most accurate samples in terms of OCR do not always yield good percentages on downstream NLP tasks. For example, *1907, vol. III, Pământ, mormânt…, Parada norocului, Băieți de fată,* all score high on OCR accuracy in our *spaCy* corpus (above 95%). Nevertheless, their accuracy scores for POS tagging and lemmatization are below their respective averages in the same corpus. The same phenomenon is not as easily discernible in *Stanza*'s operation, due to better overall performance. Still, we should point out that *Europolis, 1907, vol. III, Pământ, mormânt…* and *Vijelie,* which have all fared well on OCR accuracy (above 95%), have nevertheless scored less than the total averages on POS tagging and lemmatization.[34] However, one should take into account the fact that, in the case of *Stanza,* the difference between OCR percentages and POS and lemma percentages is at most under 7 points, in each of the texts in its associated corpus. *spaCy*, on the other hand, produces discrepancies of up to almost 18 percentage points between OCR accuracy and the accuracy of the performed NLP tasks. Generally, as Van Strien et al. (2020) have shown, if OCR accuracy percentages are above 80% (ideally 90%), the results of certain NLP tasks are not significantly inaccurate. Likewise, Strange et al. (2014, para. 52), although not particularly concerned about NLP, conclude that cleaning an OCR corpus is "desirable but not essential". Therefore, it remains to be seen whether OCR accuracy is below the mentioned thresholds at the level of the entire DMRN III archive. However, within the limits of our accuracy measurements, both off-the-shelf models tested here performed reasonably well with *Stanza* being, as we have seen, more reliable in its correlations between the quality of the OCR and the performance of its functions.

5. Limits

In the course of our experiment, we have identified certain limits to our methodology:

1. In regard to the initial sampling of the DMRN III digital archive, possibly a more representative sample of the archive should have taken into account the quality of the OCR for each individual file. Since at the level of the entire archive, this certainly influences the quality of NLP, a prior evaluation of each entry on the basis of a scale (from very good, to bad OCR) would have provided more adequate *strata* for our sampling. This, however, was beyond the possibilities of our small research team.
2. One major limitation of the present study is the absence of a *gold standard*, greatly felt in measuring the quality of the OCR of our samples. Our accuracy measurements in this sector are therefore not very trustworthy, these samples being too small for general assessments.[35] Given the magnitude of the DMRN archive and the diversity of the original print editions (and also the amount of manual labor involved), it is hard to

[34] Also, *Zaza*, which scored 99.10% accuracy in the OCR department in this corpus, acquired only 90.99% at POS tagging.
[35] Although the overall OCR accuracy percentages in both our corpora (*spaCy*'s and *Stanza*'s, respectively) are almost identical – which might suggest that the measurements taken are accurate –, we would like to point the reader to the accuracy scores for *Porți fără număr, vol. II*: 82,86% in the *spaCy* corpus, and 93,79% in *Stanza*'s. This variation can only be explained by the insufficient size of our sample and statistical chance.

ISSN: 2003-0924

believe that such a gold standard could ultimately be elaborated. Any other effort at performing NLP on this digital archive should therefore consider parcelling the texts on grounds of structural similarity, or limiting the selection to a few titles, so that accuracy assessments become feasible.

3. Upon cleaning, the hyphens employed to separate words on two different rows of texts were automatically removed. Still, some of these hyphens were certainly part of compound words, which were not excepted during the cleaning process. The number of such situations is not negligible, as Palmer (2010, p. 18) observes, quoting a study which states that up to 5% of the compound words occurred on different rows in an English language corpus.
4. Text segmentation in some cases has outputted punctuation as independent sentences (e.g. ")", "..."). This has reduced the number of tokens available for analysis in those particular cases. At any rate, we have decided not to replace the said 'sentences' with other randomly sampled sentences, since the number of tokens per sentence is rather a question of the author's style, as one can observe in Cezar Petrescu's case, where the small number of tokens is explained by the author's austere style of narration and dialogue, employing sequences of very short sentences.
5. A somewhat minor limitation, but one which might influence to a certain degree the accuracy of NLP tools employed on texts in the DMRN archive are the fairly frequent phrases or expressions in languages other than Romanian (mainly French, but also German and other languages of circulation), which were commonplace in the literary texts of the era. In theory, this limitation could be overcome by employing language identification NLP models.[36] In practice, the success of this method is highly dependent on OCR performance on multilingual texts.
6. There are, as we have already pointed out, significant discrepancies between the POS tags adopted by the Universal Dependencies framework (and employed by both *spaCy* and *Stanza*) and some questionable or even ambiguous aspects of the Romanian morphology. A more fitting evaluation should rather concentrate on the morphological analysis that these tools provide, than on POS tagging. This is nonetheless more time consuming, and beyond the possibilities of our research team.

6. Further Improvements and Conclusions

Certain aspects of the cleaning process could be improved. The most error prone step in our automatic cleaning process is the removal of the author's name and the novel's title from the header section in some of the documents in the DMRN archive. A better *RegEx*, one that matches the searched patterns even when the characters in the expressions are separated by any number of spaces (as was often the case with our sample documents), should improve cleaning in this regard. Also, employing such Python libraries as *difflib* (Python Software Foundation, n.d.) to match similar sequences of text, based on a similarity score of 0.7 or greater, definitely increases the success rate of removing those header section elements where the OCR engine misidentified some characters, as we demonstrated in a study that we ran after performing the evaluation done here (Vătavu & Morariu, 2024, p. 68).

As for the accuracy of the NLP, some improvement in this sector should theoretically be conceivable. One possibility is completely redoing the OCR of the entire DMRN archive, perhaps by training an open-source OCR engine (such as *Tesseract*) on the documents in the archive specifically. This, however, would prove very resource intensive and most probably

[36] Both *Stanza* (Stanford NLP Group, n.d.-a) and *spaCy* (Explosion, n.d.-b) are equipped for such tasks.

yield negligible returns in terms of better accuracy. Post-OCR correction should therefore be more adequate in this respect. Nevertheless, if we entirely exclude manual correction, solutions thus far have rarely proved significantly successful. Généreux and Spano (2015) have employed NLP in attempting to clean a German language corpus through both edit distance and string approximation. The results were "not so good", in the authors' own words. On the contrary, their methods in fact inserted new errors into the OCRed text. More successful in their attempts were Kim et al. (2021), who trained a transformer model on a corpus where OCR errors were manually tagged, then used this model to correct impressive amounts of literary texts. Although this method has also inserted errors into the texts themselves, the corrected errors were six times as many as those introduced.[37]

From our own observations, certain improvements in the accuracy of the NLP might be gained from specifically training the language models employed in this experiment on manually corrected samples extracted from the DMRN archive. This should yield better results in some of the instances where the orthography and spelling of the time have prevented the language models employed by us from accurately tagging and lemmatizing certain tokens (both words and punctuation). Nevertheless, these instances are not that numerous.

Our experiment, while having certain limitations, has turned out to be successful in many regards. First, our automatic cleaning method has proven quite adept at text sectioning. Although it did not eliminate altogether the need for further manual cleaning, we estimate that it certainly saves significant amounts of time-consuming manual labour. Furthermore, the cleaning script also delivers more suitable text files for downstream NLP tasks. As for our secondary objective, our tests prove that, despite the fact that the texts in our corpus are written in mid-20th century Romanian – including all the lexical, grammatical and orthographic aspects, the accuracy percentages that *Stanza* and *spaCy* have scored for POS tagging and lemmatization are satisfactory overall. However, *Stanza*'s performance at the time the research unfolded was significantly better, which makes it the more appropriate candidate for more extensive studies on the DMRN digital archive. Sentence segmentation, on the other hand, did not go as well as expected, and should probably be considered with more care and scepticism in future NLP analyses of the DMRN archive. Nonetheless, our methodology shows that new research avenues are possible and advisable, and that the gap observed by McGillivray, Poibeau and Fabo (2020) between the DH and NLP communities can be successfully bridged. We consider that our experience regarding the use of NLP tools to perform analyses on corpora specifically compiled for the digital humanities is valuable not only for studies that focus on Romanian literary texts from the 20th century, but also for any other DH studies that concentrate on non-English corpora.

References:

Amza, G. M., & Bilciurescu, A. (1938). *Vampirul* [The vampire]. Editura Librăriei Principele Mircea.

Anesiea, I. (1941). *Zmeu de mare* [Sea dragon] (vol. 1). Tip. Vremea.

Baghiu, Ș. (2018). Strong domination and subtle dispersion: A distant reading of novel translation in communist Romania (1944–1989). In M. Sass, Ș. Baghiu & V. Pojoga (Eds.), *The culture of translation in Romania* (pp. 65–87). Peter Lang.

[37] The method employed by Kim et al. also corrects the "intentional misspellings in dialogue". But, as the authors argue, "much of them may actually improve downstream NLP tasks even if the new words may be against the author's original intentions".

ISSN: 2003-0924

Baghiu, Ș., Pojoga, V., Bâlici, M., Chiorean, M., Ciorogar, A., Codină-Brenda, J., Crăciun, B., Farmatu, T., Mîrț, A., Morariu, D., Olaru, O., Rădescu, C., Savin, A., Stanislav, C., Stoica, A.-M., Strugari, N., Terian, A., Ung, S., Vancu, R., … Văsieș, A. (2021). *Muzeul digital al romanului românesc: 1933–1947* [The digital museum of the Romanian novel: 1933–1947] [Data set]. Complexul Național Muzeal ASTRA. https://revistatransilvania.ro/mdrr

Baghiu, Ș., Pojoga, V., Borza, C., Coroian-Goldiș, A., Gârdan, D., Modoc, E., Susarenco, T., Vancu, R., & Varga, D. (2019a). *Muzeul digital al romanului românesc: Secolul al XIX-lea* [The digital museum of the Romanian novel: The 19th century] [Data set]. Complexul Național Muzeal ASTRA. https://revistatransilvania.ro/mdrr

Baghiu, Ș., Pojoga, V., & Sass, M. (Eds.). (2019b). *Ruralism and literature in Romania*. Peter Lang.

Bart J. (1933). *Europolis*. Editura Adeverul.

Bâlici, M. (2018). The emergence of quantitative studies: Actual functionalities and the Romanian case. *Metacritic Journal for Comparative Studies and Theory, 4*(2), 54–71.

Bâlici, M. (2019). Studii cantitative recente în spațiul românesc. Între analiză instituțională și problema traducerilor [Recent quantitative studies in Romania: Between institutional analysis and the topic of translation]. *Transilvania*, *2*, 11–18.

Barbu Mititelu, V., Irimia, E., & Tufiș, D. (2014). CoRoLa—The Reference Corpus of Contemporary Romanian Language. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1235–1239. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/360_Paper.pdf

Barbu Mititelu, V., Irimia, E., Păiș, V., Avram, A.-M., Mitrofan, M., & Curea, E. (2020). Romanian resources in linguistic linked open data format. *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 29–40.

Barnoschi, D. V. (1942). *Stăvilare. Se aleg apele* [Floodgates. The waters clear]. Editura Națională Gh. Mecu.

Borza, C. (2019). How to populate a country: A quantitative analysis of the rural novel from Romania (1900–2000). In Ş. Baghiu, V. Pojoga, & M. Sass (Eds), *Ruralism and literature in Romania* (pp. 21–39). Peter Lang.

Busuioc, M., & Caragea, D. (2019). ROMTEXT: Flux de obținere și tratare a textelor [Romtext: Flow for electronic texts' generation and treatment]. *Studii și Cercetari Lingvistice*, *70*(1), 118–123. Scopus.

Ciorogar, A., & Modoc, E. (2019, February 8). Analiza computațională în cadrul studiilor literare românești [Computational analysis in Romanian literary studies]. *Observator Cultural*, *981*. https://www.observatorcultural.ro/articol/analiza-computationala-in-cadrul-studiilor-literare-romanesti/

Ciorogar, A., Modoc, E., Goldiș, A., Mudure, M., & Ursa, M. (2019, January). Studiile cantitative și provocările lor [Quantitative studies and their challenges]. *Cultura*, *1*(597), 33–39.

Cohen, M. (2002). *The sentimental education of the novel* (2nd. ed.). Princeton University Press.

ISSN: 2003-0924

Coroian-Goldiș, A., Gârdan, D., Morariu, D., Borza, C., Modoc, E., & Susarenco, T. (2019). Arhivele romanului românesc și posibilități de digitizare [The archives of the Romanian novel and digitization possibilities]. *Revista Transilvania*, *10*, 1–8.

Cotovu, S. (1947). Vijelie [Storm]. Editura Casa Școalelor.

Damrosch, D. (2006). World literature in a postcanonical, hypercanonical age. In H. Saussy (Ed.), *Comparative literature in an age of globalization* (pp. 43–53). Johns Hopkins University Press. https://www.academia.edu/download/38530018/WL_in_a_Postcanonical_Age.pdf

Del-Vet, A. (1936). *Două lumi* [Two worlds]. Editura Librăria Academică.

Dessila, O. (1946). *Porți fără număr* [Unnumbered gates] (vol. 2). Editura Cartea Românească.

Dinu, L. P., Niculae, V., & Sulea, M.-O. (2012). Pastiche detection based on stopword rankings: Exposing impersonators of a Romanian writer. In E. Fitzpatrick, J. Bachenko & T. Fornaciari (Eds.), *Proceedings of the Workshop on Computational Approaches to Deception Detection*, 72–77. Association for Computational Linguistics. https://aclanthology.org/W12-0411

Est, E. (1939). *Zaza*. Editura Țicu I. Eșanu.

Explosion. (n.d.-a). *Models and Languages.* https://spaCy.io/usage/models

Explosion. (n.d.-b). *Multi-language*. https://spaCy.io/models/xx.

Florescu, M. (1945). *Voluntarii* [The volunteers]: vol. 1. *Spre Spania* [Toward Spain]. Editura Scânteia.

Gârdan, D. (2018). Mapping emotions in the Romanian erotic novel of the interwar period: Canonical affect and popular sensibility. *Dacoromania Litteraria*, *5*(1), 101–114.

Gârdan, D. (2019). Interstitial spatiality in the Romanian novel of the interwar period: Mute rurality and subverted urbanity. In Ș. Baghiu, V. Pojoga & M. Sass (Eds), *Ruralism and literature in Romania* (pp. 69–80). Peter Lang.

Gârdan, D., & Modoc, E. (2020). Mapping literature through quantitative instruments: The case of current Romanian literary studies. *Interlitteraria*, *25*(1), Article 1. https://doi.org/10.12697/IL.2020.25.1.6

Gavrilă, V., Băjenaru, L., Dobre, C., & Tomescu, M. (2021). Towards the development of a Romanian lexicon for the analysis of emotions in the literary works of canonical authors. *Studies in Informatics and Control*, *30*(2), 111–120. https://doi.org/10.24846/v30i2y202110

Généreux, M., & Spano, D. (2015). NLP challenges in dealing with OCR-ed documents of derogated quality. *Workshop Proceedings 'Replicability and Reproducibility in Natural Language Processing: Adaptive Methods, Resources and Software' at IJCAI 2015*, 6. https://www.researchgate.net/publication/281112670_NLP_challenges_in_dealing_with_OCR-ed_documents_of_derogated_quality

Goldiș, A. (2014). Digital humanities – o nouă paradigmă teoretică? [Digital humanities – A new theoretical paradigm?]. *Revista Transilvania*, *12*, 1–4.

ISSN: 2003-0924

Goldiș, A., & Modoc, E. (2020). Distant Reading – o nouă paradigmă de cercetare literară? [Distant reading – A new literary research paradigm?]. *Revista Vatra*, *8–9*, 44–121.

Graur, A. (2009). *Mic tratat de ortografie* [Brief orthography treatise] (L. Groza, Ed.). Humanitas.

Hanu, B., Vlad, A., & Mitrea, A. (2018). Aspects revealing the orthography and punctuation impact in printed Romanian: A literary corpus based study. *2018 International Conference on Communications (COMM)*, 95–100. https://doi.org/10.1109/ICComm.2018.8484819

Ion, R., Irimia, E., & Barbu Mititelu, V. (2018). Ensemble Romanian dependency parsing with neural networks. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1574–1579.

Istrate, I., Milea, I., Modola, D., Pop, A., Popa, M., Sasu, A., Stan, E., Taşcu, V., & Vartic, M. (2004). *Dicționarul cronologic al romanului românesc de la origini până la 1989: DCRR* [The chronological dictionary of the Romanian novel from its origins to 1989]. Editura Academiei Române.

Kim, A., Pethe, C., Inoue, N., & Skiena, S. (2021). Cleaning dirty books: Post-OCR processing for previously scanned texts. *Findings of the Association for Computational Linguistics, Findings of ACL: EMNLP 2021*, 4217–4226. Association for Computational Linguistics.

McGillivray, B., Poibeau, T., & Fabo, P. R. (2020). Digital humanities and natural language processing: "Je t'aime... Moi non plus". *Digital Humanities Quarterly*, *14*(2). https://doi.org/10.17863/CAM.55816

Mieskes, M., & Schmunk, S. (2019). OCR quality and NLP preprocessing. In A. Axelrod, D. Yang, R. Cunha, S. Shaikh & Z. Waseem (Eds.), *Proceedings of the 2019 Workshop on Widening NLP* (pp. 102–105). Association for Computational Linguistics. https://aclanthology.org/W19-3633

Missir, I. (1937). *Fata moartă* [The dead girl]. Editura Cartea Românească.

Modoc, E. (2018). Traveling avant-gardes: The case of futurism in Romania. In Ş. Baghiu, V. Pojoga & M. Sass (Eds), *The culture of translation in Romania* (pp. 47–65). Peter Lang.

Modoc, E. (2020). *Internaţionala periferiilor: Reţeaua avangardelor din Europa Centrală şi de Est* [The international of peripheries: Avant-garde networks of East-Central Europe]. Editura Muzeul Literaturii Române.

Modoc, E., & Gârdan, D. (2020). Style at the scale of the canon: A stylometric analysis of 100 Romanian novels published between 1920 and 1940. *Metacritic Journal for Comparative Studies and Theory*, *6*(2), 48–63. https://doi.org/10.24193/mjcst.2020.10.03

Montani, I., Honnibal, M., Honnibal, M., Boyd, A., Van Landeghem, S., & Peters, H. (2023). Explosion/spaCy: V3.7.2: Fixes for APIs and requirements (Version v3.7.2) [Computer software]. *Zenodo*. https://doi.org/10.5281/zenodo.10009823

Morariu, D., Stoica, A.-M., Farmatu, T., Vancu, R., & Varga, D. (2021). Poli de producție ai romanului românesc (1933-1947): Rețele editoriale și forme de canonizare [Production centers of the Romanian novel (1933-1947): Editorial networks and forms of canonization]. *Revista Transilvania*, *9*, 35–42. https://doi.org/10.51391/trva.2021.09.05

Moretti, F. (2005). *Graphs, maps, trees: Abstract models for a literary history*. Verso.

ISSN: 2003-0924

Neagu, L. M., Dascalu, M., Trausanmatu, S., Simion, E., & Chisu, L. (2020). Automated modeling of Romanian literary trends in history using topics over time and co-occurences. *Proceedings of the 16th International Scientific Conference 'eLearning and Software for Education' Bucharest, April 30 - May 1, 2020*, 151–158. https://doi.org/10.12753/2066-026X-20-019

Nicolaescu, M., & Mihai, A. (2014). Teaching digital humanities in Romania. *CLCWeb: Comparative Literature and Culture*, *16*(5). https://doi.org/10.7771/1481-4374.2497

Nunberg, G. (1990). *The linguistics of punctuation*. Center for the Study of Language and Information.

Olaru, O. (2019). What is digital humanities and what's it doing in Romanian departments? *Revista Transilvania*, *2019*(5–6), 30–37.

Păiș, V., Ion, R., Avram, M., Mitrofan, M., & Tufiș, D. (2021). In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology*, *24*(4), 384–401.

Palmer, D. D. (2010). Text preprocessing. In N. Indurkhya & F. J. Damerau, *Handbook of natural language processing* (2nd ed., pp. 9–30). Chapman and Hall/CRC.

Patraş, R., Galleron, I., Grădinaru, C., Lionte, I., & Pascaru, L. (2019). The splendors and mist(eries) of Romanian digital literary studies: A state-of-the-art just before Horizons 2020 closes off. *Hermeneia*, *23*, 207–222.

Patraș, R., Galleron, I., Pascariu, L., Olteanu, A., Lionte, I., & Grădinaru, C. (2020). *HAIRO: corpus de romane haiduceşti (1850-1950)* [HAIRO: Corpus of hajduk novels (1850-1950)] [Data set]. https://www.nakala.fr/page/collection/11280/51a881d1

Petrescu, C. (1943). 1*907*: vol. 3. *Pământ, mormânt…* [Land, grave…]. Cugetarea – Georgescu Delafras.

Pojoga, V. (2018). A survey of poetry translations in Romanian periodicals (1990–2015). In M. Sass, Ş. Baghiu & V. Pojoga (Eds), *The culture of translation in Romania* (pp. 99–121). Peter Lang.

Pojoga, V., Baghiu, Ş., Modoc, E., Gârdan, D., & Coroian Goldiș, A. (2019). Tehnici digitale pentru analiza romanului românesc [Digital tools for the analysis of the Romanian novel]. *Revista Transilvania*, *10*, 9–16. https://revistatransilvania.ro/wp-content/uploads/2019/12/02.Vlad-Pojoga-Stefan-Baghiu-Emanuel-Modoc-Daiana-Gardan-Andreea-Coroian-Goldis.pdf

Pojoga, V., Neagu, L. M., & Dascalu, M. (2020). The character network in Liviu Rebreanu's *Ion*: A quantitative analysis of dialogue. *Metacritic Journal for Comparative Studies and Theory*, *6*(2), 23.

Popescu, M., & Dinu, L. P. (2008). Rank distance as a stylistic similarity. In D. Scott & H. Uszkoreit (Eds), *Coling 2008: Companion volume: Posters* (pp. 91–94). Coling 2008 Organizing Committee. https://aclanthology.org/C08-2023

Poppler (n.d.) [Computer software]. https://poppler.freedesktop.org/

Prinsloo, D., Taljard, E., & Goosen, M. (2022). Optical character recognition and text cleaning in the indigenous South African languages. *Stellenbosch Papers in Linguistics Plus*, *64*, 165–187. https://doi.org/10.5842/64-1-867

ISSN: 2003-0924

Python Software Foundation. (n.d.). *difflib – Helpers for computing deltas*. https://docs.python.org/3/library/difflib.html

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). https://doi.org/10.18653/v1/2020.acl-demos.14

Rebreanu, L. (1940). Amândoi [Both of them]. Editura Socec & Co.

Sass, M., Baghiu, Ş., & Pojoga, V. (2018). *The culture of translation in Romania*. Peter Lang.

Stanford NLP Group. (n.d.-a). *Language identification*. https://stanfordnlp.github.io/stanza/langid.html

Stanford NLP Group. (n.d.-b). *Model performance*. https://stanfordnlp.github.io/stanza/performance.html

Stănoiu, D. (1934). *Parada norocului* [The parade of fortune]. Editura Librăriei Universala Alcalay & Co.

Strange, C., McNamara, D., Wodak, J., & Wood, I. (2014). Mining for the meanings of a murder: The impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, *8*(1). https://digitalhumanities.org/dhq/vol/8/1/000168/000168.html

Streinul, M. (1944). *Băieți de fată* [Illegitimate sons]. Editura ABC.

Take. (1935). *Nunta Lizetii* [Lizeta's wedding]. Editura Cartea Românească.

Terian, A., Baghiu, Ș. (2025). Digitising literary heritage: Some lessons from the Digital Museum of the Romanian Novel. In T. Lähdesmäki, J. Turunen, A. Terian & R. Garcia-Bardidia (Eds.), *Engaging communities in cultural heritage* (pp. 113-127). Routledge.

Terian, A. (2019). Big numbers: A quantitative analysis of the development of the novel in Romania. *Transylvanian Review*, *XXVIII* (Suppl. 1), 55–74.

Terian, A., Farmatu, T., Borza, C., Varga, D., Văsieș, A., & Morariu, D. (2021). Genurile romanului românesc (1933-1947). O analiză cantitativă [The genres of the Romanian novel (1933-1947): A quantitative analysis]. *Transilvania*, *9*, 43–54. https://doi.org/10.51391/trva.2021.09.06

Tufiș, D., & Dan, C. (2018). A bird's-eye view of language processing projects at the Romanian Academy. In N. C. (Conference chair), K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis & T. Tokunaga (Eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Unicode. (n.d.). *Unicode normalization FAQ*. https://unicode.org/faq/normalization.html

Universal Dependecies. (n.d.-a). *Statistics of UD Romanian RRT*. https://universaldependencies.org/treebanks/ro_rrt/index.html#statistics-of-ud-romanian-rrt

Universal Dependencies. (n.d.-b). *UD for Romanian*. https://universaldependencies.org/ro/index.html

ISSN: 2003-0924

Universal Dependencies. (n.d.-c). *Universal Dependencies*. https://universaldependencies.org/

Ursa, M. (2015). Is Romanian culture ready for the digital turn?. *Metacritic Journal for Comparative Studies and Theory*, *1*, 80–97.

Van Strien, D., Beelen, K., Ardanuy, M., Hosseini, K., McGillivray, B., & Colavizza, G. (2020). Assessing the impact of OCR quality on downstream NLP tasks. *Proceedings of the 12th International Conference on Agents and Artificial Intelligence*, 484–496. https://doi.org/10.5220/0009169004840496

Vătavu, B., & Morariu, D. (2024). Enhancing usability of digital collections: Accuracy assessment and OCR post-correction of the Digital Museum of the Romanian Novel. *Transilvania*, *10*, 66–75. https://doi.org/10.51391/trva.2024.10.08

Vintilă-Rădulescu, I., Rădulescu Sala, M., & Aranghelovici, C. (2021). *Dicţionarul ortografic, ortoepic şi morfologic al limbii române: DOOM* (Ediţia a 3-a revăzută şi adăugită) [Orthographic, orthoepic and morphological dictionary of the Romanian language. Revised and updated, 3rd Edition]. Univers Enciclopedic Gold.

Xiao, R. (2010). Corpus creation. In N. Indurkhya & F. J. Damerau (Eds), *Handbook of natural language processing* (2nd ed., pp. 147–165). Chapman and Hall/CRC.

ISSN: 2003-0924