Emanuel MODOC
Sextil Puşcariu Institute of Linguistics and Literary History, The Romanian Academy
Cluj-Napoca, Romania
emanuel.modoc@academia-cj.ro

THE DIGITAL CURATION OF THE ROMANIAN INTERWAR NOVEL (1920-1940)

Recommended Citation: Modoc, Emanuel. "The Digital Curation of the Romanian Interwar Novel (1920-1940)". *Metacritic Journal for Comparative Studies and Theory* 8.1 (2022). DOI: https://doi.org/10.24193/mjcst.2022.13.07.

Abstract: In recent years, Romanian literary studies took one of its major methodological turns toward distant reading, using either or both quantitative and computational analysis. While quantitative analysis employed lexicographical instruments such as dictionaries and literary chronologies, computational analysis tried to approach the issue from a "data rich" historical perspective (Katherine Bode), while also attempting to build a digital corpus adapted to computational methods. The following paper attempts to survey the main research projects that deal with the computational analysis of Romanian literature in general and the Romanian novel in particular. The first part of the study undertakes a succinct state-of-the-art on past and ongoing endeavours concerned with digital approaches to the study of Romanian literature, their initial findings and potential. The second part will take a more theoretic approach to some of the key concepts related to data supported literary history. Finally, the last part of the study tackles the main challenges of developing a digital corpus of a local literature and the shortcomings related to this literature's "locality" in terms of computational approaches and the compatibility of the tools developed by Western research projects.
Keywords: distant reading, Digital Humanities, Romanian literature, Romanian novel, computational literary studies

Computational approaches to the study of literature have fast become one of the main tendencies in recent years to tackle complicated issues in the humanities. What

this domain brings forth is an inter- and transdisciplinary approach to the survey of national cultural phenomena by way of methods and instruments stemming from the hard sciences (information technologies, statistics etc.). One of the central objectives of computational literary studies (CLS) is to lay bare at least some of the otherwise unperceivable aspects pertaining to a given literary production, while also building up a comprehensive digital archive able to conserve its cultural heritage. Because such an approach is dependent on large quantities of information and aims to extrapolate general principles from – at the very least – a representative sample corpus, computational analysis strives to aid traditional methods of studying literature in understanding notions such as the evolution of literary forms, genre formation, style, or other formal changes.

In Romania, computational methods in literary investigations started with modest corpuses, garnered the interest of a relatively small number of researchers (most of them coming from linguistics) and privileged the novel over other literary genres. The predominant focus on the novel owes much to the genre's importance in the development of Romanian literary modernity. Even though poetry had its role in the shaping of modernist sensibility, the novel's relevance also extended in the socio-cultural domain, since as early as late nineteenth century it encompassed many extra-literary functions: as a benchmark for social and cultural changes in the nation, as a repository of national traditions and mentalities, as a method of propagating the core tenets of national identity, as a means to combat repressive political policies, and, lastly, as an instrument of social emancipation and progress.

Previous attempts to study Romanian literature from a Digital Humanities perspective can be classified according to their main research subject and the proportions of the corpuses used. Concerning the Romanian novel, three main projects were developed in recent years. The first one, Astra Data Mining: The Digital Museum of the Romanian Novel has managed to archive to an almost exhaustive extent the novelistic production published between 1845 and 1947 (Baghiu et al, Muzeul). Within this collective research project, a series of studies employing metadata analysis starting from the corpus emerged (Coroian-Goldiş, Arhivele; Terian et al, Genurile romanului XIX; Pojoga, Tehnici; Terian et al, Genurile romanului 1901-1932; Baghiu et al, Geografia romanului XIX; Baghiu et al, Geografia romanului 1901-1932; Modoc et al, Geografia romanului 1901-1932; Modoc et al, Geografia romanului 1933-1947) that tackled aspects such as subgenre development,

literary geography, digital archives or digital methods in literary analysis. Another research project, this time focusing on a specific subgenre, was led by Roxana Patraş. The Hai-Ro (*Hajduk Novels in Romania during the Long Nineteenth Century: Digital Edition and Corpus analysis Assisted by Computational Tools*) deals with the *hajduk* novels published between 1850 and 1950 and proposes the creation of a digital corpus using digital XML/ TEI markups together with semantic annotations. Hai-Ro's development owes much to the transnational COST project *European Literary Text Collection* (ELTeC), which undertook the creation of a multi-linguistic corpus of European novels published between 1850 and 1920. The creation of this pan-European corpus was designed as a necessary starting point for the development of computational instruments able to deal with all European languages included in the project, Romanian included. Finally, the last project, ARCAN (The Romanian Novel from Archive to Canon) reunites qualitative formal instruments (mainly genre theory) and computational tools (with a heavy emphasis on NLP technologies) in order to investigate the relationship between the literary canon of the period 1845-1947 and the "literary archive" of the era.

Before the appearance of these projects, however, the Romanian critical space went through a period of theoretical debate and metacritical reflection. Tracing the emergence and development of computational studies in the West, Alex Goldiş published one of the first articles concerned with Digital Humanities and quantitative studies (Goldiş 1-4). This early intervention is rightfully cautious when talking about the possibility of importing distant reading methodologies in the study of Romanian literature, as in the absence of a comprehensive digital literary archive, the type of macroanalysis proposed by Jockers is all but impossible to emulate. The following year, Mihaela Ursa asked herself if Romanian culture is ready for "the digital turn", outlining the country's cultural and institutional shortcomings and the local academe's lack of flexibility toward intermediality and interdisciplinarity (Ursa 80-97). Another important moment in this self-reflexive period is Roxana Patraş, Andreea Mironescu, Camelia Grădinariu and Emanuel Grosu's 2017 article *Can Interdisciplinarity Be Measured? A Quantitative and Qualitative Analysis of Keywords*. Here, the authors ponder the question of terminological vulnerability, stating that terms such as multi-, post-, pluri-, trans-, and interdisciplinarity have become interchangeable and, thus, superfluous, opting for Digital Humanities – which is *inherently* multi-, post-, pluri-, trans-, and interdisciplinary – as a means to

overcome this handicap (Patraş et al 17-31, see also Nicolaescu and Mihai). Later studies built upon these first reflections (Bâlici 54-71, Olaru 30-37, Gârdan and Modoc 52-65) after the first autochthonous forays into distant reading have appeared in the country.

Archive, corpus, canon. Digital Humanities in the age of curation

In one of the pamphlets (Algee-Hewitt et al) later published in the volume *Canon/ Archive*, the researchers at Stanford Literary Lab make an essential distinction between what they called "the published" (i.e., the total production of a given culture), "the archive" (i.e., what was conserved out of the total production) and "the corpus" (what made into their final sample collection). "[T]he fundamental horizon of all quantitative work" (Algee-Hewitt 2), *the published* is a literary sample that can never be completely surveyed through direct means, but can be accessible through its preserved metadata (by way of lexicographical instruments, for instance). *The archive*, on the other hand, is the sum total of literary works that was preserved in libraries, collections, and archives. This is the basis of any given *corpus*, the lowest common denominator of the three, which invariably implies a process of selection, since the archive (in this particular sense) cannot be readily accessible at all times. The difficulties surrounding the need to equalise the three types of collections are pertinently explored by the researchers at Stanford Literary Lab, who worked with a corpus of 4000 English novels published between 1750 and 1880, a relatively small number compared to the archive, let alone the published. Moreover, they also had to face with a selection bias in the corpus, with significantly more gothic novels than historical and large quantitative differences across different timeframes. Thus, in order to compensate for the bias, Stanford Literary Lab opted, in their experiment, for a sample of 674 novels (Algee-Hewitt et al 2-3). Setting aside the difficulties related to the data mining of this corpus, one of the authors' conclusion in this case study was that any given quantitative approach on a corpus of this magnitude relied heavily on the interaction between multiple institutions and research collectives, a notion that is, to this day, quite foreign for the Romanian humanist. In the case of the Romanian collection of novels accessible in digital format, the archive and the corpus present a much more generous ratio: 80% of the novels published in book form between 1845 and 1947 is currently available via the Astra Data Mining corpus,

according to the coordinators (Baghiu et al, Geografia 1933-1947 8)[1]. In principle, this solves the selection bias encountered by the Stanford Literary Lab researchers in their corpus.

The issue of bias is replaced, however, with the one regarding the relation between the corpus and the canon. Because the Romanian literary canon is highly restrictive, giving way to a "great unread" exponentially greater than any found in a culture with a larger literary production. Out of nowhere, the corpus made available by the Astra Data Mining comes with an abundance of data that require the ability to sort through the newly found knowledge in order to utilize it more effectively. For this, I propose a typological approach to the curation of the Romanian novel using the cultural common denominators of the Romanian novel (i.e. the established canon) and a temporal frame that can be used as the milestone for the birth and peak of the modern Romanian novel: 1920-1940. The year 1920 marks the publication of what is unanimously considered the first modern novel, Liviu Rebreanu's *Ion*, while 1940 marks the year of the last canonical novel published in the interwar period, Mihail Sebastian's *Accidentul*. For the period in question, authors such as Liviu Rebreanu, Camil Petrescu, Mihail Sadoveanu, G. Călinescu, Hortensia Papadat-Bengescu, Anton Holban, Mateiu Caragiale, Max Blecher, Mihail Sebastian, and Mircea Eliade are canonical references that have affixed in the literary canon a set of subgenres that can be used as the starting point for selecting a sampling corpus that is both expansive and representative for the study of the modern novel. Already, the authors mentioned above filter out a great deal of literary works that did not engage in the cultural dynamics of the period, while also selecting quite a large collection of minor novels with shared formal, thematic, or spatial dimensions. Thus, my canonical selection includes both a representation of the Romanian critical canon and a historically relevant sample of the Romanian field of production.

Following these basic principles, I will attempt to propose a selection based on the abovementioned aspects. First, a comprehensive list of *the published*, thanks to *The Chronological Dictionary of the Romanian Novel from its Origins to 1989*[2] (Istrate et al.). Then, the access to *the archive* courtesy of the many national libraries that preserved the novels. Finally, a part of *the corpus* provided by the project

[1] This percentage does not take into account serialized novels published in the literary periodicals of the time.
[2] As a shorthand, I shall name it DCRR from now on.

ASTRA Data Mining. A first selection is represented by the canon itself. More specifically, *canons*, because we have to take into consideration not only what our critical tradition deems canonical (canonization as a result of choices made by literary critics, historians, publishers, or the public), but also what canon is from a distant reading perspective. Thus, my canonical selection includes both a representation of the Romanian critical canon and a historically relevant sample for the Romanian literary production. Finally, the process of selecting minor novels following the dominant novelistic subgenres dictated by the canon. Because of this self-imposed rule, *genre fiction*, by far the literary form furthest from all notions of canonization, was omitted.

Another crucial aspect to my selection is that it presupposes an agreement about the specific subgenre classes in question. This is not as simple as it seems, since the only (somewhat) consistent source of traditional genre labelling in Romania remains DCRR[3]. The only aspect that simplifies the question of label adequacy is the general consensus between the canonical novels and their respective labels, since one will be hard pressed to find scholars that do not deem Liviu Rebreanu's *Ion* a rural novel or Camil Petrescu's *Patul lui Procust* [The Procustean Bed] a psychological one. Moreover, my selection does not tackle with the arduous task of using unsupervised methods of classifying novels coming from machine learning (see Underwood 34-67). Even if "genres, like buildings, possess distinctive features at every possible scale of analysis: mortar, bricks, and architecture" (Allison et al 8), attempts at trying to find markers in a Romanian corpus are hindered by a lack of compatible tools, as I will detail below.

Therefore, I have selected 10 subgenres[4] representative for their canonicity in Romanian literature: biographies (including autobiographies), erotic, sentimental (in Margaret Cohen's sense of the concept), family novels, parable, historical, psychological, rural, social, and war. In order to avoid a bias in representation (the social novel is dominant in the period, followed by the sentimental), I opted to limit the number of novels/ subgenre to fifty titles (in the rare cases in which this number was achieved). What remains is a very succinct illustration of my proposed corpus: 27 (auto)biographies (authored by canonical writers such as Max Blecher and Panait

[3] The next edition of DCRR is still in its editorial stages. See Borza et al 205-6.
[4] The typological particularities of the subgenres, as well as the classifications are borrowed from the research team responsible with the revised edition of DCRR, see Borza et al 205-220.

Istrati), 36 erotic novels (featuring authors such as Felix Aderca, Eugen Lovinescu, or Mircea Eliade), 6 family novels (with notable authors such as Hortensia Papadat-Bengescu), 35 historical novels (often teetering on the edge of popular fiction, the historical novel has been frequently used by canonical writers, the indisputable master of the genre is Mihail Sadoveanu), 8 parables (Sadoveanu, again, takes centre stage), 50 psychological novels (one of the most diverse in terms of canonical authorship: from Camil Petrescu and Anton Holban to Liviu Rebreanu and Hortensia Papadat-Bengescu), 27 war novels (Aderca, Papadat-Bengescu), 42 rural novels (with Liviu Rebreanu and Cezar Petrescu at the top of the canonical list), 50 sentimental novels, and 50 social novels. The three subgenres that have been capped completely dominate the novelistic production and have quite an imbalanced canon-to-archive ratio, but in turn feature a very diverse selection of authors. This shows that, at least in part, subgenres that can permeate different groups of writers (specialized, amateurs) and different groups of public (high-, middle-, and low-brow alike) have the best quantitative representation in the field. Because most of the subgenres that display a high degree of canonicity rarely reach this number, the total amount of novels in my selection amounts to 331 novels out of a total of 859 published in the period (according to DCRR). With a coverage just shy of 40% of the total production, my selection is useful in covering the part of the archive that was most engaged in the dynamic cultural field of the interwar period.

An abundance of data, a lack of tools

Following the main digitization projects mentioned above, Romanian culture has quickly gained access to an abundance of *data*. Before that, scholars relied heavily (and many still do) on lexicographical proxies in order to compensate for the lack of data with metadata. Subgenres, years of publication, publishing houses, the main cities housing most of the Romanian editorial landscape etc. were systematically used to survey the Romanian novel *in its absence*. Even now, the archive still needs work. As of yet, no project has been able to deliver a consistent enough corpus of annotated novels or a sufficiently encompassing structured dataset. It is an endeavour that takes a great deal of consistent collective effort and an even greater deal of raising awareness of the fact that with digitization comes an urgent need for collective research groups and the establishment of strong relational ties between institutions.

The limits of current efforts to employ computational analysis on a Romanian corpus are obvious: lack of encoded, machine-readable corpus, limited implementation of NLP tools for writing with a high degree of historical character, absence of markup standards for Romanian (either procedural, presentational or descriptive). One of the more basic instruments for computational textual analysis that also feature NLP support for Romanian is TXM (*Textométrie*). However, the tool is highly dependent on a stable form of language and the presence of encoded text, as it is primarily a linguistics tool, capable of lexical operations such as progressions, concordances or co-occurrences. More complex operations such as topic modelling or sentiment analysis, methods that feature unsupervised data collocation, automated semantic clustering or machine learning integration are all but impossible to implement at this time. Other methods that do not depend necessarily on NLP integration (since it relies on normalized word frequencies/ z-scores) such as stylometry (computational stylistics) show some promise (see Modoc and Gârdan 48-63)[5], but nevertheless require further research in order to confirm its validity on Romanian language corpuses.

In conclusion, the domain of Romanian literary studies is still in its very early stages when it comes to employing methods of computational analysis. It cannot be overstated how much future institutional strategies and policies will play their part in the continued establishment of Digital Humanities in the Romanian academe. While the projects discussed above may seem modest in their scope, they are indicative of a paradigm shift in the field of literary studies. It is expected that in the following few years even the shortcomings mentioned earlier will be overcome. The extent to which Digital Humanities will exact change in the autochthonous academic field will depend on the future outcomes that will have built upon these humble beginnings. Finally, it should be remembered that, numbers-wise, the current available digital archive of Romanian novels has a degree of coverage impossible to achieve in any Western culture. Perhaps for the first time in the history of literary studies,

[5] In this article, we have not used a balanced corpus, nor was it representative of the entire literary production. Representing 13% of the period's production and with a 50% presence of canonical works, our study merely intended to test the merits and/ or limits of stylometry on a literary corpus without NLP support. In some cases, authors marked as stylistically special by traditional literary historiography do end up in an isolated position within the network, while, in other cases (Camil Petrescu, Mircea Eliade or G. Călinescu), they are a great deal more closely linked to the rest of the authors. We believe that these results have the potential to reopen some debates concerning personal style using formal markers such as stop-words.

peripheral cultures have a more than privileged position in terms of being able to render and investigate exhaustively their complete literary heritage.

Acknowledgement: This work was supported by a research grant financed by the Recurring Donor's Fund at the Romanian Academy and managed by "PATRIMONIU" Foundation, project number: GAR-UM-2019-I-1.5-6: "Gestiunea digitală a romanului românesc. Digitalizarea patrimoniului de roman autohton 1920-1940" (GDRR).

References:

Algee-Hewitt, Mark et al. "Canon/ Archive. Large-scale Dynamics in the Literary Field". Stanford Literary Lab. January 2016: https://litlab.stanford.edu/LiteraryLabPamphlet11.pdf.

Allison, Sarah et al. "Quantitative Formalism: An Experiment". Stanford Literary Lab. January 2011: https://litlab.stanford.edu/LiteraryLabPamphlet1.pdf.

Baghiu, Ştefan et al. "Geografia internă a romanului românesc în secolul al XIX-lea" [The Internal Geography of the Romanian Novel in the 19th Century]. *Transilvania*, no. 10, 2019, pp. 29-43.

Baghiu, Ştefan et al. "Geografia romanului românesc (1901-1932): Străinătatea" [The Geography of the Romanian Novel (1901-1932): Spaces from Abroad]. *Transilvania*, 10, 2020, pp. 1-11.

Baghiu, Ştefan et al. *Muzeul Digital al Romanului Românesc: 1901-1932* [*The Digital Museum of the Romanian Novel: 1901-1932*]. Complexul Naţional Muzeal ASTRA, 2021: https://revistatransilvania.ro/mdrr1901-1932.

Baghiu, Ştefan et al. *Muzeul Digital al Romanului Românesc: 1933-1947* [*The Digital Museum of the Romanian Novel: 1933-1947*]. Complexul Naţional Muzeal ASTRA, 2022: https://revistatransilvania.ro/mdrr1933-1947.

Baghiu, Ştefan et al. *Muzeul Digital al Romanului Românesc: secolul al XIX-lea* [*The Digital Museum of the Romanian Novel: the Nineteenth Century*]. Complexul Naţional Muzeal ASTRA, 2020: https://revistatransilvania.ro/mdrr.

Bâlici, Mihnea. "The Emergence of Quantitative Studies. Actual Functionalities and the Romanian Case". *Metacritic Journal for Comparative Studies and Theory*, 4.2, 2018, pp. 54-71.

Borza, Cosmin et al. “Subgenurile romanului românesc. Laboratorul unei tipologii”. *Dacoromania litteraria*, vol. 4, 2020, pp. 205-220.

Burrows, John F. “Delta: A Measure for Stylistic Difference and a Guide to Likely Authorship”. *Literary & Linguistic Computing*, 17.2, 2002, pp. 267-87.

Coroian-Goldiş, Andreea et al. “The Archives of the Romanian Novel and Digitization Possibilities”. *Transilvania*, nr. 9, 2019, pp. 1-8.

Eder, Maciej et al. “Stylometry with R: A package for computational text analysis”, *R Journal*, 8.1, 2016, journal.r-pro-ject.org/archive/2016/RJ-2016-007/index.html), pp. 107-121.

Gârdan, Daiana, and Emanuel Modoc. “Mapping Literature through Quantitative Instruments. The Case of Current Romanian Literary Studies.” *Interlitteraria*, 25.1, 2020, doi.org/10.12697/IL.2020.25.1.6, pp. 52-65.

Gârdan, Daiana, and Emanuel Modoc. “Mapping Literature through Quantitative Instruments. The Case of Current Romanian Literary Studies”. *Interlitteraria*, 25.1, 2020, pp. 52-65.

Goldiş, Alex. “Digital Humanities – o nouă paradigmă teoretică?”. *Transilvania*, no. 12, 2014, pp. 1-4.

Istrate, Ion et al. *Dicţionarul cronologic al romanului românesc de la origini până în 1989* [Chronological Dictionary of the Romanian Novel from its Origins to 1989]. Editura Academiei Române, 2004.

Modoc, Emanuel et al. “Geografia romanului românesc (1901-1932): arealul national.” *Transilvania*, no. 10, 2020, pp. 12-21.

Modoc, Emanuel et al. “Geografia romanului românesc (1933- 1947): arealul naţional”. *Transilvania*, no. 9, 2021, pp. 10-17.

Modoc, Emanuel, and Daiana Gârdan. “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, 2020, pp. 48-63. DOI: https://doi.org/10.24193/mjcst.2020.10.03.

Moretti, Franco (Ed.). *Canon/ Archive*. n+1 Foundation, 2017.

Nicolaescu, Mădălina, and Adriana Mihai. “Teaching Digital Humanities in Romania”. *CLCWeb: Comparative Literature and Culture*, 16.5, 2014.

Olaru, Ovio. “What is Digital Humanities and What’s It Doing in Romanian Departments?”. *Transilvania*, no. 5-6, 2019, pp. 30-7.

Patraş, Roxana et al. “Poate fi măsurată interdisciplinaritatea? O analiză cantitativă şi calitativă a cuvintelor-cheie”. *Transilvania*, no. 10, 2017, pp. 17-31.

Patraş, Roxana et al. “The Splendors and Mist(eries) of Romanian Digital Literary Studies: a State-of-the-Art just before Horizons 2020 closes off” . *Hermeneia*, no. 23, 2019, pp. 207-222.

Patraş, Roxana. “Hayduk novels in the nineteenth-century Romanian fiction: notes on a sub-genre”. *Swedish Journal of Romanian Studies*, 2.1, 2019, https://doi.org/10.35824/sjrs.v2i1.18769.

Pojoga, Vlad et al. “Digital Tools for the Analysis of the Romanian Novel”. *Transilvania*, no. 10, 2019, pp. 9-16.

Pojoga, Vlad et al. “The Character Network in Liviu Rebreanu's Ion: A Quantitative Analysis of Dialogue”. *Metacritic Journal for Comparative Studies and Theory*, 6.2, 2020, pp. 22-46.

Terian, Andrei et al. “Genurile romanului românesc (1901-1932)” [Genres of the Romanian Novel (1901-1932)]. *Transilvania,* no. 10, 2020, pp. 53-63.

Terian, Andrei et al. “Genurile romanului românesc în secolul al XIX-lea. O analiză cantitativă.” *Transilvania*, no. 10, 2019, pp. 17-28.

Underwood, Ted. *Distant Horizons. Digital Evidence and Literary Change*. University of Chicago Press, 2019.

Ursa, Mihaela. “Is Romanian Culture Ready for the Digital Turn?”. *Metacritic Journal for Comparative Studies and Theory*, 1.1, 2015, pp. 80-97.