Roxana PATRAŞ [*], Ioana GALLERON[**]
Camelia GRĂDINARU[***], Ioana LIONTE[****]
Lucreţia PASCARU[*****]

The Splendors and Mist(eries) of Romanian Digital Literary Studies: a State-of-the-Art just before Horizons 2020 closes off [******]

Abstract: The present article is a snapshot of Digital Literary Studies (DLS) in the present-day Romanian academia, higher education curricula, and research evaluation. In the first part, the emphasis falls on the term "digital turn" and on its specific uses and extensions in humanities, as DH (digital humanities), on the one hand, and as digital literary studies/ computer literary studies (DLS/ CLS)/ computational linguistics (CL), on the other. In the second part, we zoom in the field of DLS/ CLS and analyze the way in which it has been localized, operationalized, institutionalized and understood in the Romanian academic environment and publications (DH-targeted journals, humanities journals, and cultural magazines), in higher education curricula (master/ bachelor programs of study), and in designing evaluation standards for DH/ DLS/ CLS research projects (methodologies for funding national research). In the third part, we provide a down-to-earth approach to Romanian DLS by bringing out the experience with digitization, format conversion, manual cleaning, encoding, annotation, and with various editing, quantitative analysis, and data management tools (AntConc, TXM, StyloR, Nooj, Heurist, Transkribus, Oxygen etc.), acquired throughout the implementation of Hai-Ro Project (*Hajduk Novels in Romania during the Long Nineteenth Century: digital edition and corpus analysis assisted by computational tools*).

Keywords: Digital Literary Studies, Digital Turn, digitization, data management, annotation.

[*] Senior Researcher (CS II), Institute for Interdisciplinary Research, Sciences and Humanities Research Department, "Alexandru Ioan Cuza" University of Iaşi, Romania; email: roxana.patras@uaic.ro

[**] Professor, Sorbonne Nouvelle University, LATTICE research team (UMR 8094), France; email: ioana.galleron@sorbonne-nouvelle.fr

[***] Senior Researcher (CS III), Institute for Interdisciplinary Research, Social Sciences and Humanities Research Department, "Alexandru Ioan Cuza" University of Iasi, Romania; email: camelia.gradinaru@uaic.ro

[****] Assistant Professor, PhD student, "Alexandru Ioan Cuza" University of Iaşi, "Gr. T. Popa" University of Medicine and Pharmacy, Romania; email: ilionte@yahoo.com

[*****] Assistant Researcher, PhD student, "Alexandru Ioan Cuza" University of Iaşi, Romania; email: lucrelucre9@gmail.com

[******] **Acknowledgement**: This work was supported by a grant of the Romanian Ministry of Research and Innovation, CCCDI - UEFISCDI, project number PN-III-P3-3.1-PM-RO-FR-2019-0063 / 13 BM/2019, within PNCDI III.

1. Introduction

Digital humanities (DH) are at the crossroads between two cultures: humanities and computation. This field became widespread since 1990s, but its complexity makes it hard to define even today (Schnapp and Presner 2009). It supposes the use of analogue and digital sources, a hybrid methodology, an interdisciplinary framework and a various range of technologies (databases, data analytics, linguistic analysis software, geographical and social mapping tools and so on). Moreover, DH does not mean the simple application of digital tools to already existent data, but it implies a profound level of speculative thinking, creativity and adaptation. Thus, the spirit of DH is "of experimentation along the entire work chain: theorizing and conceptualization, research, data collection, content curation, data processing, data analytics, and often open publishing (of digital corpora and collections, of virtualized experiences, of publications, and of multimedial presentations)" (Hai-Jew 2017, ix). The digital tools and software are not used only in order to extend humanities research, but also to deeply reflect on how methodologies could shape our interpretation of data. In this vein, "digital humanities projects are not simply mechanistic applications of technical knowledge, but occasions for critical self-consciousness" (Drucker and Nowviskie 2004, 432). The ground on which humanists work is fundamentally changed and an "algorithmic criticism" (Ramsey 2011) could be found at work.

The digital turn – expression seen by Mills (2010) as a "pun" on Gee's "social turn" in literacy studies (2000) – is bringing new genres, ways of editing and modelling and, in sum, new modes of knowledge. In humanities, the digital turn expanded not only the research material, but also the research questions. Thus, "data, once captured, cleaned and encoded, could be easily interrogated using simple methods but from a variety of perspectives, allowing researchers to escape disciplinary silos so their work better reflected the complexity which humanities seek to make sense of" (Cosgrave 2019, 9).

If in architecture Carpo (2017) talked about "the second digital turn", in humanities the power of changes is considered seminal. The "generative humanities" represent "a mode of practice that depends on rapid cycles of prototyping and testing, a willingness to embrace productive failure, and the realisation that any 'solutions' generated within the Digital Humanities will spawn new 'problems'" (Burdick et al. 2012, 5). The DH' effects can be perceived also in the deconstruction of the artificial divide between humanities and sciences, showing that humanists, together with scientists, are still needed to solve contemporary problems (Liu 2012; Fiormonte et al. 2015).

During the last decades, DH centers or teams have started to flourish in various higher education and research institutions around the globe, and this new discipline – or maybe an interdisciplinary field, or a set of research methods, as some may say[1] – became rapidly in fashion, as proven amongst others by the rapid renaming of other endeavors on the same model: researchers can nowadays engage into "medical humanities", "spatial humanities", "climate change humanities" and so on (Schreibmann, Siemens and Unsworth, 2016). Borne by the cultural and political accent fallen on the complexity of present-day life challenges, the metaphor of the "crossing"[2] has proven thus to be a fertile one, inviting humanities researchers to open up to other approaches, epistemic frameworks and tools, so as to produce new ideas and insights.

2. Digital Humanities in Romania

Without any pretense on rendering an exhaustive overview, we could spot 2014 as a moment of emergence for Romanian DH studies, more notably, in the field of literary studies. Initially, DH occurred in articles authored by Romanian scholars as a hazy concept that called for either polemic action or theoretical conjectures. Over the last 5 years, DH has legitimized itself as a theoretical paradigm by appealing to field-related *glossae* rather than data-driven research; however, curriculum initiatives and evaluation standards (for funding national projects) caught the new buzz in the air, launching Master's Programs (University of Bucharest) and designing a special domain for DH-related projects in the last UEFISCDI calls (https://uefiscdi.gov.ro/p1-dezvoltarea-sistemului-national-de-cd).

In what follows, we are sketching a timeline for the development of DH research in Romania and propose a categorization based on the publication types we could identify. This reveals not only the way in which DH is theoretically negotiated and conceptually managed locally, but also the fact that this umbrella term is usually associated with research on metadata which does not ground on results yielded by actual digital tools. Accordingly, we have identified the following types of DH articles:

a. DH articles that showcase the premises and/or results of research by emphasizing the general lines and the work-in-progress particularities;

b. DH articles that take inventory and discuss in a general note the advancements of the field itself (software tools, computational adjustments, etc.) but without trying them on Romanian texts;

c. DH articles that use the term "digital humanities" either as a conceptual counterpart or as a taxonomical correspondent in order to advent an emerging field and thereafter to jumpstart a more extensive debate/analysis/research that makes use of related concepts such as "distant reading",

"quantitative studies", "big data", "macroanalysis", "digital literature", "intermediality" etc.

d. DH articles that try to provide conceptual/theoretical/paradigmatic insight into the perils and benefits in using the dichotomy between "digital" and ("traditional" or "national") humanities; more often than not, these have polemical aspect, for the stakes regard a paradigmatic shift in a field of study known for its proverbial resistance to change; optionally, the academic and institutional validation is also aimed at.

e. DH articles that review volumes/ pieces of research undertaken in the field of (global) Digital Humanities.

Before we proceed with detailing some of the articles' content, it is worthwhile noting that "digital humanities" is secondary (as frequency of usage) to Moretti's concept of "distant reading". Comments on Moretti's research as well as on the gracious trinity "distant reading"-"quantitative analysis"-"world literature" are, subsequently, a sort of Trojan horse that may also encapsulate some hints on DH.

The categories provided hereafter do not cover the entire spectrum but may organize a critical view on the field's recent developments. Similarly, there are articles that easily fit into more than one category as well as articles (reviews, for instance) that overlap the type.

2.1. *DH articles*

There are some publications, other than those adjacent or directly related to literary or language studies, that have endeavored to disseminate the results of DH research. These articles are written by computing specialists, computer science scholars, programmers and IT engineers who take a direct interest in the field but for whom data is always data, thus nothing more than binary computing. Their articles deal with the technical intricacies of computational work, which is actually the basis of DH studies. However, it falls within our area of interest also to take into account those "midway" publications and articles that reflect the synergy between the previously exclusive subject fields of CS and language/ literature studies, therefore between the (innovating) Digital and the (traditionalist) Humanities. Such articles can be found in *Studia UBB Digitalia*, a journal affiliated with the Transylvania Digital Humanities Centre that, since its creation in 2017, has issued four thematic volumes mentioned next in a chronological order: *Digitising the Humanities* (Moldovan and Schuster 2017), *Computing History. Eastern European Scholars* (Moldovan and Schuster 2017), *Digital Economy and Humanities* (Stanca 2018), and *Digital Classics and Ancient History* (Varga 2018).

Given the publication's transdisiplinary ambitions, one should expect a fair degree of thematic variability which *Studia UBB Digitalia* does not fail to deliver: challenges of TEI encoding and manuscript transcription (Bleier

2017, 9-25), software development and CS altogether such as eXist DB or Saxon/C in PHP (Schwaderer 2017, 100-111), articles that deal with the digital dissemination of scientific and editorial practice in terms of publishing platforms and specific software such as HTML5, CSS3, JavaScript, LaTeX (Constantinescu 2017, 42-56), metadata of photographic objects (Das Gupta 2017, 57-74). Nevertheless, few are the articles that inquire into Romanian-language corpora and databases.

2.2. *Digital Humanities and its greedy siblings: distant reading, computational analysis, quantitative studies*

Most of the articles pertaining to this section are very recent undertakings of (chiefly) literary scholars, doctoral students or graduates who conducted their researches in the wake of the new paradigms of quantitative analysis and distant reading, researches making use of a network of conceptual tools that their authors unequivocally relate to the field of DH. Analysis mainly consists of a conceptual inquiry of the subject matter: the theoretical framework of distant reading, state-of-the-art considerations concerning quantitative analysis, macroanalysis, big data, world literature, all of them envisaged under the umbrella-term "Digital Humanities". Listing research difficulties (the faults and oversights of the existing corpora, the lack of appropriate technological means for digitising texts, the lack of expertise in conducting DH studies and in establishing DH institutionally) also has a "flanerie" aspect as long as the applied part of this research misses from the argument.

Several such articles can be found in the 2019 thematic supplement of the *Transylvanian Review* titled "Romanian Literature in the Digital Age" as well as in the collective work *New Paradigms in Contemporary Romanian Literary Studies (I)* initiated by the same publication and set out, as deduced from the coordinators' introductory statement, to "get a better picture on contemporary literary research" (Baghiu and Modoc 2019, 13-16). Usually graphs are appended in order to show – and not just tell – that research on metadata is done seriously, that categories are clear-cut, and that all possible in-between items have been properly put in the right boxes. Ştefan Baghiu's *The French Novel in Translation. A Distant Reading for Romania during Communism (1944-1989)* is a nice attempt at connecting world literature studies, quantitative analysis, and polysystem studies with a research on metadata provided by the *Dictionary of the Translated Romanian Novel* (DRRT 2005). Try as we might, we could not find an indication of the tools that have been used in creating the database behind the graphs illustrating "The General Timeline for the Translation of Novels in Communist Romania", "Translations of Novels from Western Countries (1944-1989)", and "Scattered Approach to Renditions of French Novels in Romania (1944-1989)" (Baghiu 2019, 88-89),

which support the periodization of translations from French (novels) during Communism.

Relying on a type of inquisitorial attitude (one has to torture one's metadata till it tells the truth), Andrei Terian talks about *Big Numbers. A Quantitative Analysis of the Development of the Novel in Romania* and arrives to conclusions by counting original and translated novels. The survey applies various types of instruments – ILO ("index of literary originality") and ILA ("index of literary autonomy") – in order to define four major periods of the Romanian novel (Terian 2019, 59). There are, however, some points that do not result clearly, for instance, the error rate in establishing the ILO/ ILA, what is the acceptable error margin for this survey, as well as the scholar's acceptation of the term "big data/ numbers". As we all probably know, the BNF, the Gutenberg, the Google books databases make big data, whereas around 2000 Romanian novels and translations do not.

Claire Clivaz makes an extremely interesting analysis of the occurrence and institutionalization of two concurrent French equivalents for the term digital humanities in her article titled "Lost in translation?": "Whilst the collective *laudatio* of the corporeal aspect of "humanités digitales" is well founded, it is nevertheless surprising that only a few scholars have noticed the return of the outmoded French word *humanités*" (Clivaz 2017, 31). Clivaz's remark should also open a discussion about the proper translation of DH in Romanian.

2.3. *Digital Humanities as a think tank*

The articles that advent DH as a paradigmatic shift in the field of Romanian literary studies may have a secondary discursive component related to the implementation aspects. They usually discuss the emergence of DH field in terms of conflict with regard to the already established *humaniores* and to more classical forms of hermeneutics. Obviously, this is boosted by the dichotomy "close reading" versus "distant reading" and by prophecies on its implications in the future on a larger scale of local/ regional/ global literary history and theory. Here and there, challenges in terms of research facilities are mentioned too.

One of the earliest articles about this topic is Alex Goldiş's "Digital Humanities – o nouă paradigmă teoretică?", which proposes "a survey on a pilot discipline" (Goldiş 2014, 1). It inevitably departs from the theoretical apparatus of Franco Moretti and Matthew L. Jockers, all the while discussing about the new way of looking at literature through the telescope of distant reading, macroanalysis and quantitative studies. Commenting on seminal texts such as *Macroanalysis. Digital Methods and Literary History*, *Distant Reading*, *Mimesis or The Rise of the Novel*, Goldiş discusses the fundamental

shifts that classical hermeneutics will probably undergo once confronted with the revolutionary DH methods.

Skimming over some aspects of computational analysis, *Analiza computațională în cadrul studiilor literare românești* provides the general readership with a very brief overview of the CA 16204, *Distant Reading for European Literary History*. The article fashions itself in terms of an alarm signal on the current obstacles of "computing" Romanian literature. Among such obstacles, the authors list the lack of collaboration among Romanian researchers (language/ literature and, respectively, computer sciences) and the resistance of Romanian literary critics to new approaches. Even if the authors fail to indicate the correct link to the project's documentation on *github* and to the project's site (https://www.distant-reading.net/) and even if they do not seem to have an idea of the design or status of the Romanian collection, the European network of literary scholars, the multilingual literary corpus ELTeC and some basic tools (oXygen, TXM, Stylo, Gephi, Palladio) — not necessarily the most appropriate for lesser resourced languages such as Romanian — are fairly mentioned (Ciorogar and Modoc 2019).

Organizational, financial and technological issues are discussed in studies such as "Teaching Digital Humanities in Romania" (Nicolaescu and Mihai 2014), "Is Romanian Culture ready for the digital turn?" (Ursa 2015), "Challenges in setting up a digital humanities center in Romania" (Moldovan and Puşcariu 2017) or "What is Digital Humanities and What's it doing in Romanian Departments?" (Olaru 2019). Mădălina Nicolaescu and Adriana Mihai present a digital initiative of University of Bucharest, which consists in creating a collection of digitized translations of Shakespeare's works. The authors suggest including digital literature "as the last chapter in courses of literary history" (Nicolaescu and Mihai 2014, 3), but they are not clear whether this new type of literature should be addressed with methods specific to traditional "literary history" or should they also be studied with digital methods.

The difficulties in setting up a DH center in Romania (2017) are brought about by Corina Moldovan and Voica Puşcariu. Mihaela Ursa's article instead launches a Mephistophelian question: *Is Romanian Culture ready for the digital turn?*. Giving a very exact diagnosis, Ursa remarks that DH advent occurred in a moment when the Romanian culture and implicitly Romanian studies have not done with old feuds. For the last century, the aesthetic principle has dominated Romanian studies and the verdicts of excellence bestowed on literary works. The conflict between research practices based on "individual authority" and those based on "collective authority," that is, the scholars' preference for individual research rather than for team-based approaches, is another drawback for the future of DH (Ursa 2015, 86). In other words, Romanian researchers tend to prefer to be lone wolves because they are always after a quick hit and a clear prestige.

"What is Digital Humanities and What's it doing in Romanian Departments?" outlines two possible scenarios in which the researcher is the main character: 1. The researcher does not have access to a digitized corpus; 2. The researcher has full access to the metadata by using various programs or browsers such as Python, Jupyter Notebook, Zotero, Palladio (Olaru 2019). To have or not to have, this is thus the question... And beyond the somehow trivial manner of putting things, the article suggests that Romanian DLS researchers do not have other than metadata.

2.4. *DH read by literary reviewers*

Book reviews constitute another way of approaching the topic of Digital Humanities in the Romanian academic environment. We could trace two of such endeavors, the first one authored by Alex Ciorogar (2015), and the second, by Alex Goldiş (2017). Referring to *Digital Humanities and the Study of Intermediality in Comparative Cultural Studies* and to *Bestseller Code*, reviewers speak about "the new theoretical trends and their shifting away from textuality, focusing instead on the vast opportunities opened up by the new materiality of digital production, distribution, and consumption" (Ciorogar 2015, 226).

In a nutshell, DH's emergence in the Romanian academic discourse is streamlined mainly via literary studies and fashions itself from a discursive-polemical-theoretical angle rather than as an actual field of study. Truth is that recent developments of DH come with a high cost for those who decide to undertake research projects that involve data analysis. Most certainly, engaging in DH is not a profitable career choice, considering the amount of unrewarded preparatory work it asks for. To put it in a simple way, much effort and patience appears to be needed before being able to start producing interesting results, to such an extent that some may wonder if the entire endeavor is worthwhile, and if we are not finally moving mountains to give birth just to a small mouse. In this respect, Nan Z Da's article "The Computational Case against Computational Literary Studies" (Nan Z Da 2019) has raised some issues that, since its publication at the beginning of 2019, have been intensely debated upon.

3. Difficulties in practice: the *Hai-Ro* project

In what follows we will list some of the difficulties a DH researcher is confronted with, taking as a case study the French-Romanian project *Hai-Ro*. We will start with a short presentation of Hai-Ro[3], whose idea came about when two enthusiastic members of the COST action *Distant Reading for European Literary History* [4] started an experimental collaboration on TEI encoding and validating a small set of hajduk novels selected for inclusion

in the Romanian collection of the ELTeC[5]. The project addresses the scarcity of DLS resources designed for Romanian language and literature by creating a literary corpus of hajduk novels (1850-1950) TEI-XML conformant and including semantic annotations. One of our dearest principles is to promote a fair use of data, thus to make our corpus available as an Open access resource for approaching the Romanian literary tradition (genres, periods, canonization mechanisms, etc.) with quantitative tools and methods.

We had a vista of our misery, only when we found out that suboptimal OCR output on non-standardized Romanian and on cheap-paper 19th century prints made us spend an average of 40 hours on cleaning manually only 100 pages. This toilsome preparation of files, the workarounds related to digitization, as well as an "on-the-go" style of learning about and experimenting with new tools brought us to several hot-button issues that might be summarized as follows:

a. the prominence of NLP approaches in Romanian research environment lead to a quasi-grammarian manner of dealing with texts; according to Chomsky-Schützenberger hierarchy (Silberztein 2013, 1-13), a grammar can turn interesting results only if it proves to be "context-sensitive";

b. the libraries' politics of digitization, chiefly oriented toward Romanian press and toward historical archives and sources, lead to a narrow range of literary resources, thus to a random or trivial (read "canonically-driven") literary sampling in the already compiled Romanian corpora (Tufiş 2018; Barbu Mititelu et al. 2017): CoRoLa[6], ROMBAC, ROCO[7], SWARA, BABEL[8];

c. building balanced *literary corpora* has always come last in the line of priorities because the national literary tradition – leaving aside the murky aisles of this term, by "national" we simply mean "language-based" – has some old and new battles to fight;

d. blame it on typically Romanian imitation/ adaptation/ "forms-without-content"/ cultural emergence or not, the Romanian academia has always been prone to put the cart before the horse, especially if the cart is a rattling palanquin such as Franco Moretti's theory on "distant reading." As a matter of fact, it was only when we jumped out of the splendid palanquin that we bitterly discovered we had no (literary) data to experiment on. At this point of discussion, we attest that we are thoroughly aware of the danger in using "literary" before binary-computed data. In any case, we prefer to take this risk rather than meshing endless *glossae* on metadata provided by lexicons, national bibliographies, library catalogues, and literature dictionaries.

Surely, some digitization projects – ex-*dacoromanica*, currently called Biblioteca Digitală a Bucureştilor[9], or the digital library of UBB[10], for instance – have yielded useful resources. Nevertheless, in the case of project-oriented research questions such as ours, we could avail of neither

scans nor interoperable formats. Realizing that we are under the ground-level of any serious quantitative pursuit (that is, clean files, preferably XML), we had no other option but to draw a side-project agenda, partly composed of militant must-do-s, and partly, of naïve wishful thinking.

First things first, let us have a glimpse on mundane matters, such as *software costs* and Eastern-European *research practices related to investment in software products*. Many of us have probably noticed that for the XML format, *Oxygen 21.1* excels other open-source software, albeit a good mark should be given to *jEdit* (except for its option on *toggle line wrap*, which is quite difficult to track). While with Xpath 2.0 functions and operators there is always enough room to experience and learn, we might readily add that spellcheck in *Oxygen* looks as miserable as in any other editor as long as we keep on using the Classic Romanian Dictionary Pack[11], which relies on comma below characters and not cedilla, thus does not support diachronic and non-standardized varieties of spelling.

Similarly, whoever has tried the new version *Abby 15 Corporate* would consider that it performs better on Romanian than other free options or than OCRs provided by scanner installation kits[12]. In any case, on Romanian 19th-century non-standardized language and on cheap popular prints, which are always delivering mind-blowing UTF-8Y code and curly cedilla for glyphs such as "ş" and "ţ", *Abby* produces not entirely messy editables (see **Figure 1**) but does a sort of default normalization in the sense that "é", "ê", "ĕ" are read as "e", ĭ is read as "i" or "l", "ó" is read as "o", while "ḍ" is read as "d" or simply not available for choice in UTF-8. By the way, if one browses through the "character map" in *Oxygen*, "ḍ" and "Ḍ" must be searched in Latin Extended Additional as in Office Word. We have not tried yet OCR4all[13], designed by the researchers of University of Würzburg (Reuil et al 2019), but it promises interesting results.

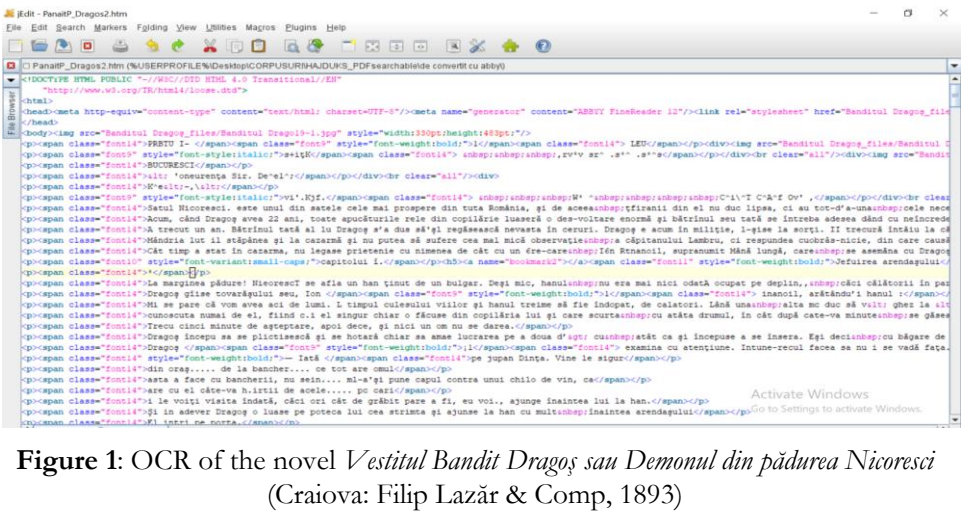

Figure 1: OCR of the novel *Vestitul Bandit Dragoş sau Demonul din pădurea Nicoresci* (Craiova: Filip Lazăr & Comp, 1893)

Anyhow, with a price over 600 $ per user license – and we have not added here the bank transfers, VAT and top-up costs related to the distribution chain – these tools are practically inaccessible to larger Romanian research teams because the entire budget of a medium research project (say, up to 100, 000 EUR) would be spent on purchasing software. There is always the Eastern European approach to research practices, which usually leads to devising genuine solutions and workarounds that are paid dearly for in terms of work-life (un)balance. Luckily enough, Hai-Ro is a bilateral French-Romanian project that could tap in resources and tools for French DH research, so all project members received a key for using Oxygen during implementation.

It was only after a *manu propria* scanning and OCR processing that we could take a closer look at texts, contents and at their particularities. Obviously, our list of 40 candidates, our theoretical assumptions on Romanian popular fiction and on novel genres, our idea of annotating spatial entities in hajduk novels had been crystallized long before the project kicked off. The only problem was that fine-grained theoretical insights would not help much when some of our novels – *Aldo şi Aminta*, Costache Boerescu's novel, for instance – stubbornly refused digitization. Illustrating those wonderful convolutions of the Romanian transition alphabet (Cazimir 2006), which obviously resulted in unacceptable OCR (see **Figure 2**), they needed a special treatment. As everything else in our corpus! The solution to this issue was Transkribus[14], a platform which enables users to train handwriting recognition models and, in problematic cases, to treat prints as manuscripts and letter fonts as handwriting. After a careful cleaning of pages, stretching of baselines, checking of text regions, line-by-line Layout analysis, and finally transliteration, we were able to train a HTR which performed pretty well on transition prints (see **Figure 3**).

The next step was an experimental use of several tools designed for quantitative analysis and data management (StyloR, AntConc, TXM, HumaNum, Heurist, and just recently Nooj), some of them working marvels on resourced languages such as English and French. While they could not be turned into palatable scientific prose (see **Figures 4, 5, 6** for several experiments with Stylo, TXM), the results of these experiments formed a pattern of prerequisites for Romanian DLS: if texts are not properly cleaned, then tokenization is not relevant; and when you manage to have a good-enough tokenization on 19th century texts, this is not enough because a highly inflected and non-standardized language such as modern Romanian will probably need a good lemmatizer. And all this is necessary just to be able to count properly; to be sure that lemmas are on the right ranks, and bring forth what some critics already deem as "a bag of words".

Then you would probably like to have some morphological and syntactic information added to strings of words resulted from queries, thus POS would be a nice feature, especially for diachronic varieties of spelling. But this will only be possible if training unsupervised tagging will work properly on our texts. However, consistent POS tagging needs normalization to a certain extent (that is, consistent principles of editing), so we return to the old feud between (original) form preservers and content divers.

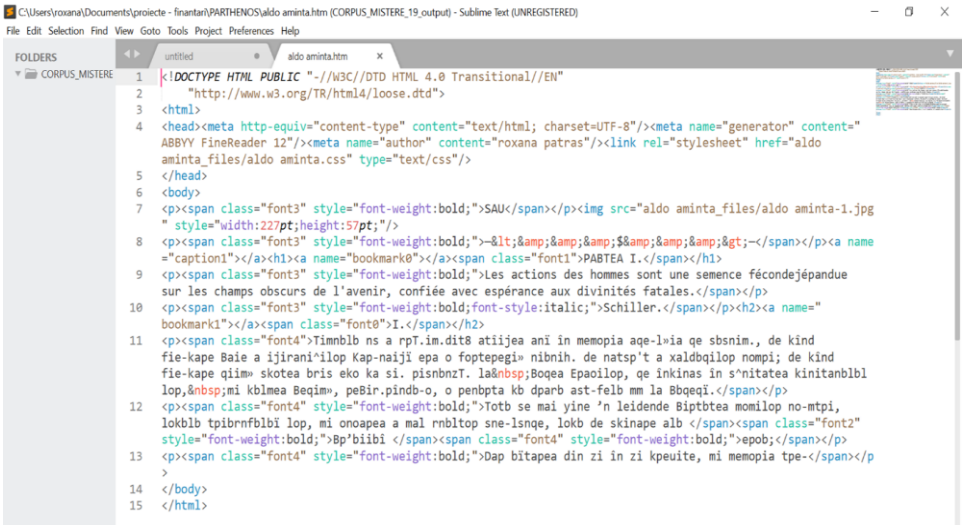

Figure 2: OCR of the novel *Aldo şi Aminta*
(Bucureşti: Tip. Bisericească din Sf. Mitropolie, 1855)

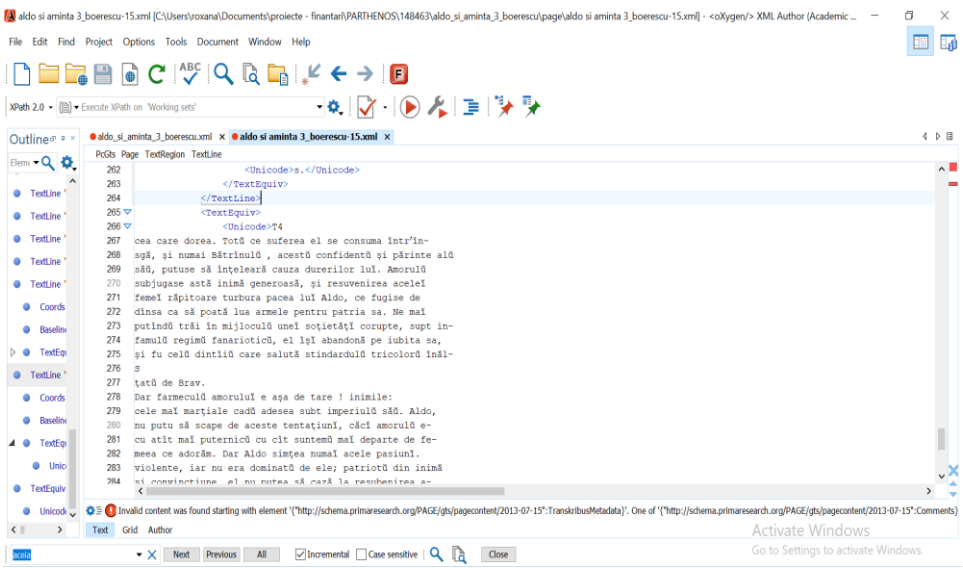

Figure 3: *Aldo si Aminta*, page 15,
automated transliteration with HTR trained by Transkribus

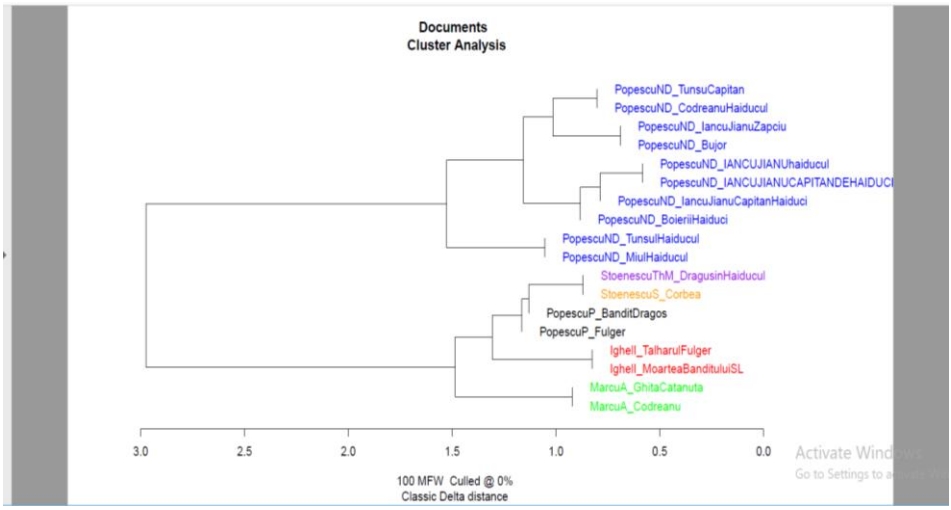

Figure 4: Experiments with immitators of N.D. Popescu (StyloR)

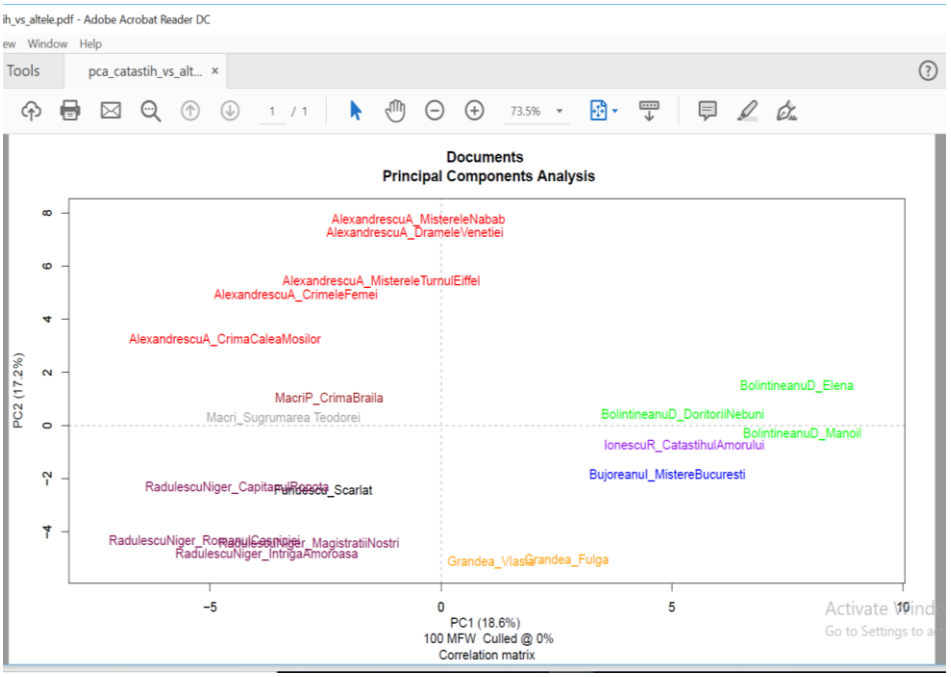

Figure 5: Experiments on authorship attribution
(Radu Ionescu's *Catastuhul amorului*)

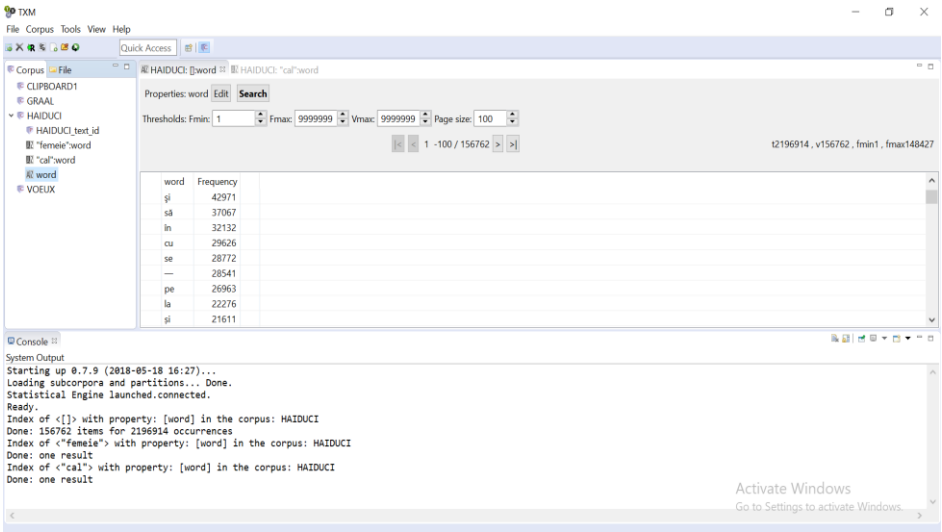

Figure 6: Experiments with TXM (suboptimal tokenization due to lack of lemmatizers and commabelow characters and cedilla characters, e.g. the conjunction "şi")

The detailed process of creating a literary corpus in a lesser resourced language such as Romanian resembles any story of raise and fall, of enthusiasm and demotivation, of splendors and miseries. Yet, while still counting on words and not only counting words, the only thing we are left with is to read Balzac's inspiring title in a playful way. If only the misery of being always late was outstretched, then the intriguing mysteries and mist(eries) of pioneer research might just yield the real meaning of splendor.

Notes

[1] For this discussion, see Schreibmann, Siemens and Unsworth, 2016. Significantly, ADHO, the international umbrella association for digital humanities, does not provide a definition for DH either on its "About" page or on other pages.

[2] "Digital Humanities is the discipline born from the intersection between humanities scholarship and computational technologies. It aims at investigating how digital methodologies can be used to enhance research in disciplines such a History, Literature, Languages, Art History, Music, Cultural Studies and many others. Digital Humanities holds a very strong practical component as it includes the concrete creation of digital resources for the study of specific disciplines." (see Pierazzo 2011).

[3] https://proiectulbrancusihairo.wordpress.com/

[4] https://www.distant-reading.net/

[5] https://www.distant-reading.net/eltec/

[6] Inappropriate for complex queries http://corola.racai.ro/

[7] Available only on ELRA http://catalogue.elra.info/en-us/repository/browse, and only under license for non-ELRA members

[8] Both of them speech corpora https://speech.utcluj.ro/swarasc/, with the important detail that BABEL is also an ELRA product, thus under license

[9] https://www.bibmet.ro/biblioteca-digitala-bucurestilor/
[10] http://dspace.bcucluj.ro/
[11] https://extensions.openoffice.org/
[12] We experienced Canon's Iris Scan Desk 5 Pro https://www.irislink.com/EN-RO/c1956/IRIScan-Desk--5-Pro---Desktop-camera-scanner.aspx
[13] https://github.com/OCR4all/OCR4all
[14] https://transkribus.eu/Transkribus/

References

Baghiu, Ştefan and Emanuel Modoc. 2019. "New Paradigms in Contemporary Romanian Literary Studies (I)". *Transilvania* (5-6): 13-16.

Baghiu, Ştefan. 2019. "The French Novel in Translation: A Distant Reading for Romania during Communism (1944–1989)". *Transylvanian Review* 28, Supplement 1: 83-100.

Barbu Mititelu, V. et al. 2017. *Corpus of Contemporary Romanian. Architecture, Annotation Levels and Analysis Tools*. In *Lingvistică românească, lingvistică Romanică*, edited by Helga Bogdan Oprea et al, 13-20. Bucureşti: Universităţii din Bucureşti.

Bleier, Roman. 2017. "Digital documentary editing of St Patrick's epistles. Linking the manuscript witnesses to the canonical text". *Studia UBB Digitalia* 62(1): 9-25.

Burdick, Anne, Drucker, Johanna, Lunenfeld, Peter, Presner, Todd and Schnapp, Jeffrey (eds.). 2012. *Digital humanities*. Cambridge, Massachusetts: MIT Press.

Carpo, M. 2017. *The second digital turn: design beyond intelligence*. Cambridge: MIT Press.

Cazimir, Ştefan. 2006. *Alfabetul de tranziţie*. Ediţia a 2-a. Bucureşti: Humanitas.

Ciorogar, Alex, and Emanuel Modoc. 2019. "Analiza computaţională în cadrul studiilor literare româneşti". *Observator Cultural* 981 (August).

Ciorogar, Alex. 2015. "Review of *Digital Humanities and the Study of Intermediality in Comparative Cultural Studies* by Steven Tötösy de Zepetnek (ed.)". *Metacritic Journal for Comparative Studies and Theory* 1(1) (October): 266-275.

Clivaz, Claire. 2017. "Lost in translation? The odyssey of 'digital humanities' in French". *Studia UBB Digitalia* 62(1): 26-41.

Constantinescu, Nicolaie. 2017. "Stretching the boundaries of publishing: The Open Web Platform and the alternatives". *Studia UBB Digitalia* 62(1): 42-56.

Cosgrave, Mike. 2019. "Digital humanities methods as a gateway to inter and transdisciplinarity". *Global Intellectual History*, 1-10, https://doi.org/10.1080/23801883.2019.1657639

Da, Nan Z. 2019. "The Computational Case against Computational Literary Studies". *Critical Inquiry* 45 (Spring): 601-639.

Das Gupta, Vinayak. 2017. "Albums in the attic. An investigation of photographic metadata". *Studia UBB Digitalia* 62(1): 57-74.

Dicţionarul cronologic al romanului românesc. 2003. Vol. 1. Bucharest: Editura Academiei Române.

Dicţionarul romanului românesc tradus. 2005. Vol. 1. Bucharest: Editura Academiei Române.

Dindelegan, Mariana. 2018. "Digital and Coding Literacy for School Students". *Studia UBB Digitalia* 63(1): 55-68.

Drucker, J., and Nowviskie, B. 2004. "Speculative computing: Aesthetic provocations in humanities computing", in *A Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens and John Unsworth, 431-447. Oxford: Blackwell.

Fiormonte, D., Numerico, T., Tomasi, F., Schmidt, D., Ferguson, C., and Rockwell, G. 2015. *The digital humanist: A critical inquiry*. New York: Punctum books.

Gee, J. (Ed.). 2000. *The New Literacy Studies: From "socially situated" to the work of the social*. London: Routledge.

Goldiș, Alex. 2014. "Digital Humanities – o nouă paradigmă teoretică?". *Transilvania* 12: 1-4.
Goldiș, Alex. 2017. "Ce romane mai citesc computerele. Review of *The Bestseller Code* by Jodie Archer and Matthew L. Jockers". *Vatra* no. 5-6.
Mills, K. A. 2010. "A review of the 'digital turn' in the new literacy studies". *Review of educational research* 80(2): 246-271.
Moldovan, Corina, and Viorica Puşcariu. 2017. "Challenges in setting up a digital humanities center in Romania". *Studia Philologia* LXII(1): 247-256.
Nicolaescu, Mădălina, and Adriana Mihai. 2014. "Teaching Digital Humanities in Romania". *CLCWeb: Comparative Literature and Culture* 16(5): 1-2.
Olaru, Ovio. 2019. "What is Digital Humanities and What's It Doing in Romanian Departments?". *Transilvania* 5-6: 30-37.
Pierazzo, Elena. 2011. *Digital Humanities: a definition*, https://epierazzo.blogspot.com/2011/01/digital-humanities-definition.html.
Ramsay, Stephen. 2011. *Reading Machines: Toward and Algorithmic Criticism*. Urbana, Chicago, and Springfield: University of Illinois Press.
Reul, Christian et al. 2019. "OCR4all-An Open-Source Tool Providing a (Semi)Automatic OCR Workflow for Historical Printings". Submitted to *Applied Sciences*: 1-54, https://www.researchgate.net/publication/335737763_OCR4all_-_An_Open-Source_Tool_Providing_a_Semi-Automatic_OCR_Workflow_for_Historical_Printings
Schnapp, J. and Presner, P. 2009. *Digital Humanities Manifesto 2.0*, http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf
Schreibman, Susan, Siemens, Ray, Unsworth, John (eds). 2016. *A New Companion to Digital Humanities*. 2nd Edition. Oxford: Wiley-Blackwell.
Schwaderer, Christian . 2017. "eXist DB or Saxon/C in PHP. A comparison between two approaches for XSLT 2.0 based websites". *Studia UBB Digitalia* 62(1): 100-111.
Silberztein, Max. 2013. "NooJ Computational Devices". In *Formalising Natural Languages with NooJ,* edited by Svetla Koeva, Slim Mesfar and Max Silberztein, 1-13. Newcastle: Cambridge Scholars.
Terian, Andrei. 2019. "Big Numbers: A Quantitative Analysis of the Development of the Novel in Romania". *Transylvanian Review* XXVIII, *Supplement* 1: 55-71.
Tufiș, D. 2018. *CoRoLa Primul corpus computațional de referință pentru limba română contemporană. Market Watch* 205: 28-29.
Ursa, Mihaela. 2015. "Is Romanian Culture ready for the digital turn?". *Metacritic Journal for Comparative Studies and Theory* 1(1): 85-97.

Web:
https://epierazzo.blogspot.com/2011/01/digital-humanities-definition.html
https://transkribus.eu/Transkribus/
https://github.com/OCR4all/OCR4all
https://www.irislink.com/EN-RO/c1956/IRIScan-Desk--5-Pro---Desktop-camera-scanner.aspx
https://extensions.openoffice.org/
http://dspace.bcucluj.ro/
https://www.bibmet.ro/biblioteca-digitala-bucurestilor/
http://catalogue.elra.info/en-us/repository/browse
https://speech.utcluj.ro/swarasc/
https://proiectulbrancusihairo.wordpress.com/
https://www.distant-reading.net/https://www.distant-reading.net/eltec/
https://uefiscdi.gov.ro/p1-dezvoltarea-sistemului-national-de-cd