Roxana Patraş, Loredana Cuzmici (editors), *Tools, Methods, and Solutions for the Exploration of Romanian Corpora*
© 2024 Institutul European Iaşi, pentru prezenta ediţie

INSTITUTUL EUROPEAN
Iaşi, Str. Băluşescu nr. 2, et. I, O P 6, CP 1309
euroedit@hotmail.com.; www.euroinst.ro

Descrierea CIP a Bibliotecii Naţionale a României

Tools, methods, and solutions for the exploration of Romanian corpora /
ed.: Roxana Patraş, Loredana Cuzmici. - Iaşi : Institutul European, 2024
Conţine bibliografie
ISBN 978-606-24-0395-9

I. Patraş, Roxana (ed.)
II. Cuzmici, Loredana (ed.)

811.135.1

All rights reserved. No part of this work covered by the copyrights hereon may be reproduced or copied in any form or by any means – graphic, electronic, or mechanical, including photocopying, recording, taping – without written permission of publisher.

PRINTED IN ROMANIA

Roxana Patraș, Loredana Cuzmici
(editors)

TOOLS, METHODS, AND SOLUTIONS FOR THE EXPLORATION OF ROMANIAN CORPORA

INSTITUTUL EUROPEAN
2024

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS–UEFISCDI, project number PN-III-P1-1.1-TE-2019-0127, within PNCDI III.

The revision and editing of this collection was performed by BolderWords translation & revision and was supported by the Altissia Chair in Digital Cultures and Ethics held by Chris Tanasescu at Université catholique de Louvain.

Table of Contents

INTRODUCTORY NOTES 7

Tactics: your DH is not my DH 9
Roxana PATRAȘ

CREATING TOOLS FOR ROMANIAN CORPORA 23

RELATE: a modern processing platform for Romanian language 25
Vasile PĂIȘ, Radu ION, Andrei-Marius AVRAM, Maria MITROFAN, Dan TUFIȘ

Tools and Resources for Digitizing Historical Romanian Documents 53
Victoria BOBICEV, Tudor BUMBU,Ludmila MALAHOV, Alexandru COLESNICOV, Svetlana COJOCARU, Liudmila BURTSEVA, Cătălina MĂRĂNDUC

Digitalization of Romanian Lexicography 85
Elena Isabelle TAMBA

ADAPTING TOOLS FOR ROMANIAN CORPORA 97

Postdigital Creativity: modeling poetry translation with multiplexes, neural networks, and large language models 99
Raluca TANASESCU, Chris TANASESCU

Challenges of Digitization and Digitalization for the Study of Writing: from archival survey to process analysis 125
Georgeta CÎȘLARU

Towards a Digital Corpus-Based Method for Assessing Language Level in EFL Student Writing: a case study on Romanian undergraduate literary analyses 139
Alexandru ORAVIȚAN, Mădălina CHITEZ

DIGITIZING ROMANIAN HERITAGE 151

A Computer-Assisted Analysis of Eminescu's *Doină* 153
Ioana GALLERON

Mapping Eastern European Political Discourse and Inequalities through Public Monuments: a digital cartography crowdsourcing project 175
Voica PUȘCAȘIU

The Title of the Feuilleton Novel: a model of semantic annotation 193
Lucreția PASCARIU

Exercises in Literary Geography: where hajduks roam in the second half of the nineteenth century 227
Alexandra OLTEANU

DIGITUS DEI EST HIC! **251**

Proxy Structures in Computational Analyses of Text Corpora **253**
Laura PRICOP

The Many Faces of Virtual: charting the family resemblance network for conceptual understanding **267**
Matei Alexandru STOENESCU

CARTOGRAPHIES OF THE ROMANIAN NOVEL: THEN AND NOW **297**

The Recontextualization of a Genre **299**
Loredana CUZMICI

Notes on the Contributors **321**

INTRODUCTORY NOTES

Tactics: your DH is not my DH

Roxana PATRAȘ

The present book should have been titled *Tactics*. Inspired by recent contributions to the state of the field (Rieder and Röhle 2012, Malazita 2021), we use Michel de Certeau's term in order to enhance the challenging situation that researchers of *less(er)-resourced languages* or *under-resourced languages* tend to be placed in, chiefly when their contributions are weighed up against the *big-tent* metaphor that has been the most impactful of all DH brands (Terras 2011, Svensson 2016). According to de Certeau, *tactics* is defined as a cluster of everyday practices devised by people or, if extrapolation is accepted, by any *realia* endowed with the ability to mobilize small-scale, lesser resources in order to move and evolve in a vast territory. Tactical navigation would thus refer not only to geographic workarounds—the shortcuts that walkers take in order to adapt to large urban environments—but also to a special navigation through the knowledge landscape in general. In a context marked by head-whirling concepts such as turbo-capitalism, datafication, infoxication, algorithmic civilization, virtual reality, big data, digital turn, technocratic culture, instrumentalist episteme, anglobalism, and so forth, it is crucial and paramount to raise awareness on the tactics that researchers of *less(er)-resourced languages*, "the DH neighbouring tribes" (Bode and Longley Arthur 2014) apply in order to enhance particularities or, as one of the European Language Resource Association [ELRA] reports states, to bring to the fore such tactics that are used simply to help them survive the digital world (Report of Special Interest Group on Under-resourced Languages [SIGUL], 2019, 2020).

The very same attributes *less(er)-resourced* and *under-resourced* also call for a deeper examination, as this might unveil and thus reveal, for instance, the origins of some particular terminology that, more often than not, have been embraced uncritically, like any other stereotype. Between 1982 and 2010, the European Bureau

for Lesser-Used Languages [EBLUL] put in circulation various reports regarding languages (mostly European) that, from a demographic and geographic perspective, were considered *lesser-used* and thus endangered. EBLUL and the European Charter for Regional or Minority Languages [ECRML], that ensued the bureau's activity, were meant to represent and speak on behalf of Europe's 50 million speakers using *regional* and *minority idioms* (*non-territorial languages* included), yet without providing clear criteria or definitions for the question under debate (Bernoville 2020). Since the classification highly depended on national states, the list of languages "protected under the Charter" reflected various policies and approaches to minority issues as well as an interwoven, very complex landscape in which many of the European idioms were tagged as both majority and minority languages. Romanian and its dialects, for instance, appeared as a minority language, and it was thus subjected to future safeguarding manoeuvers, in Bosnia and Herzegovina, Croatia, Hungary, Serbia and Ukraine. Already entangled by various historical factors, the picture of ethnic minorities' language(s) would have looked even more opaquely "thick" if ECRML had also considered the minorities and diasporas developed as a consequence of massive labor migration waves.

The release of ECRML (1992) to a certain degree coincided with the establishment of the European Language Resources Association [ELRA] in 1995 as well as with the first boost of language technologies [LT] (Păiș and Tufiș 2022). This association aimed to create a catalogue of language resources (*see* ELRA catalogue at https://catalogue.elra.info/en-us/), in addition to supporting special interest groups like *Workgroup on Less-Resourced Languages* (2016) and *Special Interest Group on Under-resourced Languages* (2017). Apparently, the terminological emphasis changed from *less(er)-used* to *lesser/ under-resourced*, that is, from debatable minority considerations to more objective and seemingly universal *technicities*. However, the new understanding of language resourcefulness as a series of scalable technicist affordances—from *use(d)* to *resource(d)*—was not apt to either sort out the majority-minority conundrum or address the pejorative keys implied in prefixes such as "less" or "under". As a consequence, in the program of the last SIGUL conference (2024), one might have come across a rich assortment of *less(er)-resourced languages*, many of them non-European: Turkish, Portuguese, Cicipu,

Effutu, Mocho', Upper Napo Kichwa, Toratán, Ulwa, Uzbek, Icelandic, Bahasa Sug, Chavacano, Filipino, Iranun, Maguindanaon, Meranaw, Sinama, Teduray, and Yakan, Occitan, Irish, Luxembourgish, *Romanian*, Bavarian, Ligurian, Croatian, Serbian, Bosnian, Montenegrin, Danish, Alsatian, Poitevin-Santongeais, Nepali, Armenian, Arbëresh, Persian (Farsi), Filipino, Cebuano, Kapampangan, Hiligaynon, Ilokano, Bikolano, Waray, and Tausug, Guarani Mbya, Somali, Basque, te reo Māori, ʻōlelo Hawaiʻi (Maite, Sakti, and Soria 2024).

In further support of the reasonable arguments summarized above and contrary to positive discrimination trends that have recently emerged in Romanian academia, the present volume proves that Romanian is anything but an *under-resourced language*. In fact, recent contributions point out that two of the top challenges that Romanian DH projects have to grapple with are firstly, the infrastructure, and secondly, the digitization policies (Schöch *et al.* 2021, Chitez, Rogobete, and Foitoș 2019, Patraș 2022). While *(digital) infrastructure* has been perceived until very recently as too general, static and nonhuman (McGrail, Nieves, and Senier 2021)—because it might designate a very diverse range of things, from services, hubs, and resources to epistemic scaffolding, unconscious institutional designs, and even "infrastructural imaginaries" (Parks 2015 *qtd. in* Lach and Pressman 2021)—the problem of digitization policies also revolves around national strategies that, going back to Michel de Certeau's reasoning, tend to manifest in a hegemonic and sluggish manner (1984). The tension between the hegemony of a research system, scarred across by the ploughs of Communist ideology and by the idleness of post-Communist transition, on the one hand, and the various tactics devised by independent DH researchers, ad-lib DH hubs, and spontaneous DH initiatives, on the other, might well enable us to contextualize the comparison trap described in an excellent report on the current status of Romanian language technologies:

> Considering the neighboring EU member states in the region, the number of Romanian corpora is the lowest... Romanian lexical and conceptual resources represent approximately 10% of the available English resources, while grammar and language models represent 20%. Considering neighboring EU member states, Romanian resources of this kind are near the region's average… In spite of the

> reduced number of available language resources [that is, *corpora*], applications for different natural language processing tasks were constructed for Romanian language. This includes lemmatization, part-of-speech tagging, dependency parsing, named entity recognition, syllabification, automatic speech recognition, text-to-speech, automatic translation, punctuation restoration, terminology annotation, text classification. Following the ELE project, over 100 tools and services for Romanian language were identified. This number represents 15% of the available English tools. Looking at language tools from neighboring EU member states, we find a similar number for Bulgarian and slightly higher for Hungarian tools (Păiș and Tufiș 2022, 7–8).

In a nutshell, attributes such as *under* and *less(er)* should, in all reality, be applied with a pinch of salt, chiefly when a zoomed-in picture of technology support across (European) languages shows that everything depends on the terms and grounds of comparison. For instance, among the 11,500 metadata records on tools/services listed in the European Language Grid (ELG) Catalogue, only English, French, German and Spanish fall within good-LT-support and moderate-LT-support groups, while the majority of other analyzed languages are "fragmentarily" supported. From the viewpoint of *use* (usage in terms of demographics)-*resource* (resourcefulness) balance, which involves, as we have already mentioned, a series of scalable technicist affordances, *Figure 1: Overall state of technology support for selected European languages* (Păiș and Tufiș 2022, 14) is even more informative because it enhances, within the fragmentary supported languages, 5 sub-groups, among which Finnish, Swedish and Dutch stand out as technically backed-up despite reduced demographics: a. Finnish, Italian, Portuguese and Dutch; b. Swedish and Polish; c. Catalan, Greek, Hungarian, Romanian, Estonian, Danish, Norwegian, Lithuanian, Czech, Latvian; d. Bulgarian, Slovak, Slovenian, Croatian, Basque, Welsh, Galician, Icelandic; e. Serbian, Maltese, Irish, Luxembourgish, and Bosnian.

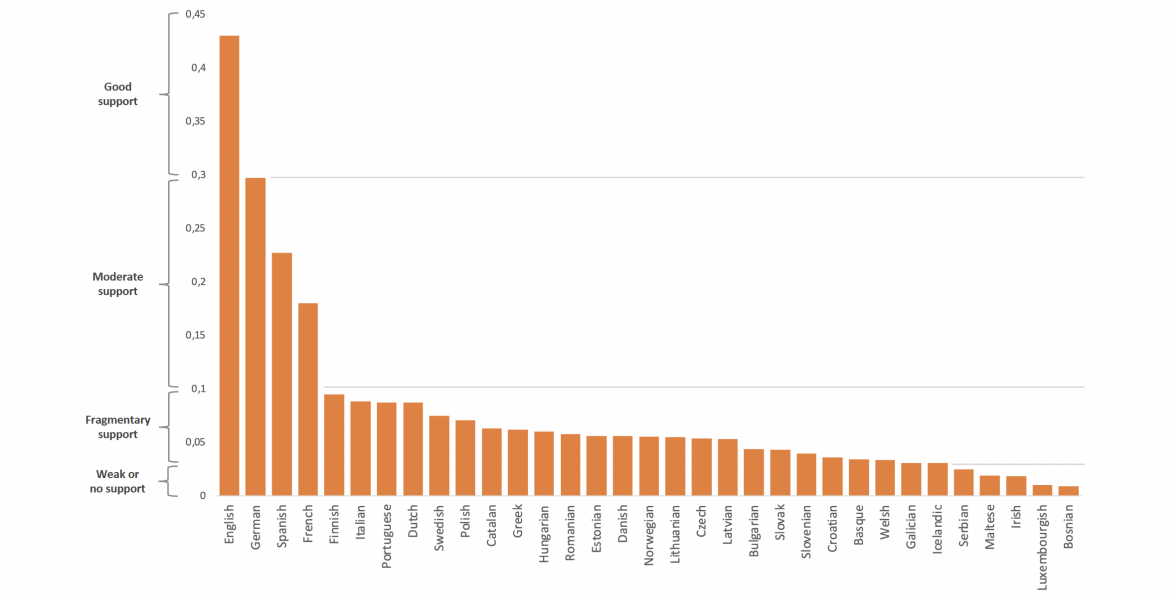


Figure 1: Overall state of technology support for selected European languages (2022)

Figure 1: Reproduced from *European Language Equality (ELE). Report on the Romanian Language*, 2022. *Image credit Vasile Păiș and Dan Tufiș*

Despite lucid epistemic calibration on ground-truth data (e. g. ELG catalogue), this kind of insights might also carve out a logic of triangulation among "alternative peripheries," which involves contrasting and negotiating the self-images of *peripheral languages* against three other cultural aspects: "the coerced and desired centres of their 'dual dependency' and the alternative periphery" (Ștefănescu 2022, 71). By surpassing the conventional postcolonialist critiques and drawing from recent input on the analogy postcolonialism-postcommunism (Chari and Verdery 2009), a transcolonial view on alternative peripheries and on alternative peripherial languages might also elucidate the majority-minority conundrum, detailed above, as well as the terms that have marked the approach, and consequently the interest, on languages usually researched and supported by the *big-tent* DH community (Weingart and Eichmann-Kalwara 2017). Indeed, if the emergence of DH as a research field has also built upon universalist and even utopian promises such as the Cold War confidence in apolitical technical expertise (Rieder and Röhle 2012, McCarty 2014, Risam 2018, Malazita 2021), then one must take into account the residual spectrality of research traditions too. The so-called *zombie concepts* (Braunstein and Warren 2021) favour all kinds of matches at the DH peripheries, for

instance the "elegiac" pairing between the computational impulse and the commemorative impulse (see also Birkerts 1994), the quixotic mélange in the formula "minimal computing" (Risam and Gil 2022) or the impetuously assertive disjunction "Your DH is not my DH" (Ray Murray 2015).

In an article entitled "Where is the ML in DH? And Where is the DH in ML? The Relationship between Modern Languages and Digital Humanities, and an Argument for a Critical DHML", Thea Pitman and Claire Taylor avouch that "it is time for all the various and different versions of DH to compare and contrast their situatedness globally, hopefully eschewing the construction of any sense of centre/periphery or 'one true DH' as the exercise unfolds" (2017). This means acknowledging the linguistically-heterogeneous landscape of global DH and endorsing multilingualism—at least as a desideratum for alternative methods, tools and concepts—in some of the most esteemed DH events and publications (see, for instance, the special issue of "Cultural Analytics" devoted to the Arab DH). However, the openness to multilingualism and diversity does not even start to grease the giant, long-established (and some would say outmoded) wheels of global DH, which are still deeply entrenched, as postcolonial critics from Tara McPherson (2012) to Roopika Risam (2018) prove, by the colonial and neocolonial influences that have molded the digital cultural record so far, or, in short, by the systemic processes of canon replication.

Assisted by data from the formerly cited reports, we may add (while hoping to conduct proper research on this matter sometime in the near future) that canon replication secretly communicates with LT replication and dissemination in the DH peripheries. Beyond all talks about digital infrastructure and policies, that is, about *strategies*, technological scalarity is applicable only to groups of good- and moderately-supported languages and partially accounts for the uneven developments in terms of text processing (e.g. part-of-speech tagging, syntactic parsing), information extraction and retrieval (e. g. search and information mining), translation technologies (e.g. machine translation, computer-aided translation), natural language generation (e.g. text summarisation, simplification), speech processing (e.g. speech synthesis, speech recognition), image/video processing (e.g. facial expression recognition), human-computer interaction (e.g. tools for conversational systems).

However, as already revealed, the fetishization of technology and the compulsive focus on a digital preservation and tool-making paradigm do not allow for the problematization of digital disparities (Fernández L'Hoeste and Rodrigues 2020), for the nuanced approach on the locus of enunciation about DH concepts (Gallego Cuiñas 2024) as well as about situated DH tools, methods and solutions, which are the focal points of the present volume.

In an attempt to address (and hopefully redress) technology fetishization—as well as a whole spectrum of emotions, from awe and disapproval to submission and optimism—that hollow out at the DH peripheries and in the emergent DH communities, the book's first two sections [*Creating tools for Romanian Corpora* and *Adapting Tools for Romanian Corpora*] follow a seamless and logical sequence all the way through to the last two sections [*Digitizing Romanian Heritage* and *Digitus Dei est Hic!*] by bridging *tools*, on the one hand, and *ethos*, on the other. When not straightforwardly protectionist, the European legaleeze on *less(er)-resourced languages* suggests that research stock should chiefly be invested in the creation of dedicated tools for endangered idioms as well as for a series of patrimonial elements enumerated in the UNESCO convention of 2003. However, as Vasile Păiș *et al.*, Victoria Bobicev *et al.* and Elena Isabelle Tamba attest, the integration of firm and already-tested solutions represents a cost-efficient and a human-empowering tactics. Addressing the fragmentation in the LT landscape and in Romanian academia, these contributions fully prove that instead of competing with one another, research projects and platforms (including NLP features such as annotation, transcription or digitization) should be designed to be interoperable and to interact with each other in order to create eco-systemical synergies. Bringing to the fore significant contributions to open access facilities, both citizen science and private initiatives are also touched upon from a critical angle, by reaffirming the FAIR principles, as well as by pleading for scholarly validation for all freely available LT resources (e.g. applications such as *https://dexonline.ro* and datasets created for various purposes).

The second section assembles contributions that share the perspective on corpora and archives as living, thus complex, realities—imaginable as *multiplexes* or as *multilayer networks*—that should thus be managed and modelled by keeping pace

with the ever-changing digital knowledge infrastructure (see also Bode 2018). Surmounting divides and embracing the concept of post-digitality, Raluca Tanasescu and Chris Tanasescu aptly point out that, while tools and code lines might belong to a largely-frequented set, it is the methods and approaches that could bring to the fore the wildly creative part of DH research. For instance, the relationship between the operation of translation and computer code reveals its fascinating versatility because it is rooted in the fact that, at its core, translating human-readable code into machine-executable instructions is a form of interpretation. Even if not targeted as such, the matters of translability and of foreign language proficiency are also at the core of Alexandru Oraviţan and Mădălina Chitez's contribution, which describes a new method for assessing language levels in EFL student writing by using ROGER as a benchmark parallel corpus. Reflecting on the changes that the process of writing undergoes in the digital environment, Georgeta Cîşlaru's chapter *Challenges of Digitization and Digitalization for the Study of Writing: from archival survey to process analysis* introduces, through means of keylog data, a very ingenious approach to the textual genetics, as well as a new take on the long-established profiling of content creators (writers, for instance) as either *discoverers* or *executors*:

> Compared to handwritten or typewritten data, texts produced by means of digital tools are formally and qualitatively different, and offer different research perspectives. Drafts become evanescent and not archival, and revisions leave no visible physical trace.

Section three *Digitizing Romanian Heritage* includes four use-cases, illustrative for the manner in which heritage texts and patrimonial corpora should be explored and promoted. Ioana Galleron's chapter draws from a larger initiative that aims at the prototypization of the first digital edition in Romania. The substantive results on the versions' melodic evolution and thematic patterns, both of them strikingly submerging or, on the contrary, raising up to the final version of the poem *Doină*, support the claim that the process of collation-comparison between the official scholarly edition and the collection of facsimiled manuscripts will be the corner stone of *Digital Mihai Eminescu*, if such project will ever catch the eye of

research strategists in Romania. While Ioana Galleron abstracts a poem's convolutions to a map, the other contributions also rely on different kinds of mapping: mapping the secret lives and conflicting ideologies of public monuments in Cluj-Napoca (Voica Puşcaşiu), mapping the patterns in titling feuilleton novels (Lucreția Pascariu), and mapping the intersections of fictional spaces and real places ventured through by one of the most iconic Romanian heroes, the hajduk (Alexandra Olteanu). Occasioned and supported by various DH projects implemented at the Digital Humanities Laboratory (Institute of Interdisciplinary Research, "Alexandru Ioan Cuza" University of Iași), these chapters are also wrapped up in the *epilogue* signed by Loredana Cuzmici who, in line with Hirsch (2012), contributes a detailed analysis on the status of digital literary corpora and on their possible uses for pedagogical purposes.

Under the facetiously ominous title *Digitus Dei est Hic!*, the last section juxtaposes two philosophical approaches to the stunning complexity of concepts such as digitality, computationality, and virtuality. While their methods seem rather dissimilar—Matei Alexandru Stoenescu explores amplification and Laura Pricop explores abstraction—these contributions provide further proof of a falsely-oppositional paradigm that has recently been discussed by Katherine Bode in "What's the Matter with Computational Literary Studies" (2023). In fact, the reader of this collection will perceive how the same rhizomatic concepts and practices could be addressed in different ways: by compiling an expansive corpus of networked semantic traits in order to empirically test Wittgenstein's concept of "family resemblances" or by identifying, in recent computational literary studies, the nodal importance of arguments by proxy in harnessing the conceptual complexities of literature and in steering away from instrumentalist shaming.

Following the notion of tactics, the chapters in this volume also critically inquire if digital humanities constitute "an opportunistic, strategic, and instrumental version of academic imperialism" (L'Hoeste and Rodrigues 2020, 14) or whether they have the potential to become a counterhegemonic practice against various conceits on lesser-resourced languages and regions (Thompson Klein 2015, Galleron and Idmhand 2020). We are grateful to our friends Chris Tanasescu, Raluca Tănăsescu, Ioana Galleron, Georgeta Cîșlaru, and Mădălina Chitez, whose mature

situatedness in both the global DH and Romanian DH communities has opened a new path for reflection on diasporic modes and alternative peripheries. We also extend our gratitude to the other collegues who are active members of DH hubs around Romania and whose gentle feedback, in all circumstances, has made our own tactics much sleeker and more entertaining.

References

Bernoville, Gabrielle. "Europe's Forgotten Words. The Case of the European Regional and Minority Languages." In *La Regionist. A Regions of Europe Society Journal*, Vol. 2, January–February, 2020, 3–8.

Birkerts, Sven. *The Gutenberg Elegies: The Fate of Reading in an Electronic Age*. London: Faber and Faber, 1994.

Bode, Katherine. "What's the Matter with Computational Literary Studies." *Critical Inquiry* 49, no. 4 (2023): 507–529.

Bode, Katherine. *A World of Fiction Digital Collections and the Future of Literary History*. Ann Arbor: University of Michigan Press, 2018.

Bode, Katherine and Paul Longley Arthur. "Collecting ourselves." In Advancing Digital Humanities. Research, Methods, Theories, edited by Paul Longley Arthur and Katherine Bode, 1–12. London: Palgrave Macmillan, 2015.

Braunstein, Laura R., and Michelle R. Warren. "Zombies in the Library Stacks." In *People, Practice, Power. Digital Humanities outside the Center*, edited by Anne McGrail, Angel David Nieves, and Siobhan Senier, 70–83. Minneapolis–London: University of Minnesota Press, 2021.

Chari, Sharad, and Katherine Verdery. "Thinking between the Posts: Postcolonialism, Postsocialism, and Ethnography after the Cold War." *Comparative Studies in Society and History* 51, no. 1 (2009): 6–34.

Chitez, Mădălina, Roxana Rogobete, and Alexandru Foitoș. "Digital Humanities as an Incentive for Digitalisation Strategies in Eastern European HEIs: A Case Study of Romania." In *European Higher Education Area: Challenges for a New Decade*, edited by A. Curaj, Ligia Deca, Remus Pricopie, 545–565. https://doi.org/10.1007/978-3-030-56316-5_34.

De Certeau, Michel. *The Practice of Everyday Life*. Berkley, Los Angeles, California: University of California Press, 1984.

ELRA Catalogue. https://catalogue.elra.info/en-us/. Accessed in May 2024.

European Charter for Regional or Minority Languages. https://rm.coe.int/1680695175. Accessed in May 2024.

Fernández L'Hoeste, Héctor, and Juan Carlos Rodríguez. "Introduction: In and Out of Digital Humanities; Nations, Networks, and Practices in Latinx America." In *Digital Humanities in Latin America. Reframing Media, Technology, and Culture in Latin/o America*, edited by Héctor Fernández L'Hoeste and Juan Carlos Rodríguez, 1–38. Gainesville: University of Florida Press, 2020.

Gallego Cuiñas, Ana. "Literature Seen Through Big Data and Artificial Intelligence: Key Concepts and Critical Challenges." In *Humanities and Big Data in Ibero-America*, 25–42. Berlin: De Gruyter, 2024.

Galleron, Ioana, and Fatiha Idmhand. "Why Go from Texts to Data, or The Digital Humanities as A Critique of the Humanities." *Word and Text. A Journal of Literary Studies and Linguistics* 10 (2020): 53–69.

Hirsch, Brett D. (ed.) *Digital Humanities Pedagogy. Practices, Principles and Politics*. Cambridge: Open Book Publishers. https://www.openbookpublishers.com/books/ 10.11647/ obp.0024.

Lach, Pamella R., and Jessica Pressman. "Digital Infrastructures: People, Place, and Passion—a Case Study of San Diego State University." In *People, Practice, Power. Digital Humanities outside the Center*, edited by Anne McGrail, Angel David Nieves, and Siobhan Senier, 189–202. Minneapolis–London: University of Minnesota Press, 2021.

Malazita, James. "Epistemic Infrastructure, the Instrumental Turn, and the Digital Humanities." In *People, Practice, Power. Digital Humanities outside the Center*, edited by Anne McGrail, Angel David Nieves, and Siobhan Senier, 3–24. Minneapolis-London: University of Minnesota Press, 2021.

McCarty, Willard. "Getting There from Here: Remembering the Future or Digital Humanities." *Literary and Linguistic Computing* 29, no. 3 (2014): 283–306.

McGrail, Anne, Angel David Nieves, and Siobhan Senier. "Introduction." In People, Practice, Power. Digital Humanities Outside the Center, edited by Anne McGrail, Angel David Nieves, and Siobhan Senier, VII-XXII. Minneapolis-London: University of Minnesota Press, 2021.

McPherson, Tara. "Why Are the Digital Humanities So White? or, Thinking the Histories of Race and Computation." In *Debates in the Digital Humanities*, edited by M. K. Gold, 139–60. Minneapolis: University of Minnesota Press, 2012.

Melero, Maite, Sakriani Sakti, and Claudia Soria, eds. "Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024." 2024. https://aclanthology.org/2024.sigul-1.pdf.

Păiș, Vasile, and Dan Tufiș. *European Language Equality (ELE). Report on the Romanian Language*. 2022. https://european-language-equality.eu/wp-content/uploads/2022/03/ ELE ___Deliverable_D1_29__Language_Report_Romanian_.pdf. Accessed in June 2024.

Patraș, Roxana. "Digital Romania: A Novelty Effect?" *Hermeneia* 29 (2022): 127–148.

Pitman, Thea, and Claire Taylor. "Where is the ML in DH? And Where is the DH in ML? The Relationship between Modern Languages and Digital Humanities, and an Argument for a Critical DHML." *DHQ* 11, no. 1 (2017).

Ray Murray, Padmini. "Locating the Digital Humanities in India: Internationalisation, Globalisation, and Localisation." Plenary presentation at Digital Diversity 2015, University of Alberta, Edmonton, Canada, May 9, 2015.

Report of Special Interest Group on Under-resourced Languages [SIGUL]. 2019-2020. https://www.elra.info/media/filer_public/2020/11/03/2019-20.pdf.

Rieder, Bernhard, and Theo Röhle. "Digital Methods: Five Challenges." In *Understanding Digital Humanities*, edited by David M. Berry, 67–85. London: Palgrave Macmillan, 2012.

Risam, Roopika, and Alex Gil. "Introduction: The Questions of Minimal Computing." *DHQ: Digital Humanities Quarterly* 16, no. 2 (2022).

Risam, Roopika. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Evanston: Northwestern University Press, 2018.

Schöch, Christof, *et al*. "Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives." *Modern Languages Open* (2021). DOI: http://doi.org/10.3828/mlo.v0i0.364.

Svensson, Patrik. *Big Digital Humanities. Imagining a Meeting Place for the Humanities and the Digital*. Ann Arbor, Mich: University of Michigan Press, 2016.

Ștefănescu, Bogdan. "The Complicated Selves of Transcolonialism: The Triangulation of Identities in the Alternative Peripheries of Global Post/Colonialism". *Metacritic Journal for Comparative Studies and Theory* 8.1 (2022). DOI: https://doi.org/10.24193/mjcst.2022.13.04.

Terras, Melissa. "Peering inside the Big Tent: Digital Humanities and the Crisis of Inclusion." Melissa Terras' Blog, July 26, 2011. peering-inside-big-tent-digital.html.

Thompson Klein, Julie. *Interdisciplining Digital Humanities: Boundary Work in an Emergent Field*. Ann Arbor: University of Michigan Press, 2015.

Weingart, S. B., & Eichmann-Kalwara, N. "What's Under the Big Tent?: A Study of ADHO Conference Abstracts." *Digital Studies/le Champ Numérique* 7, no. 1 (2017): 6. DOI: http://doi.org/10.16995/dscn.284.

CREATING TOOLS FOR ROMANIAN CORPORA

RELATE: a modern processing platform for Romanian language

Vasile PĂIȘ, Radu ION,
Andrei-Marius AVRAM,
Maria MITROFAN, Dan TUFIȘ

1. Introduction

Language Technology (LT) platforms provide services for the analysis and production of written or spoken language, making use of artificial intelligence (AI) methods to do so. Furthermore, these platforms have been developed either to showcase new technologies or to function as powerful tools for processing large amounts of data, in the form of offline corpora or online requests.

One of the principal goals of the 1st International Workshop on Language Technology Platforms (IWLTP 2020) was to address fragmentation in the language technology landscape. As the workshop organizers noted, instead of competing with one another, platforms should be designed to be interoperable and to interact with each other to create synergies towards a productive LT ecosystem (Rehm *et al.* 2020a). Interoperability is usually achieved by providing input and output in standardized formats, specific to the tasks at hand. Interaction between systems involves making the functionalities of one system available for use by another system.

In this paper, we present the RELATE platform (Păiș *et al.* 2020), a modular state-of-the-art platform that is used for processing the Romanian language developed at the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy (for short RACAI in English, or ICIA in the Romanian language). Integrating resources and technologies developed in our institute as well as those developed by partner institutions, RELATE is actively being used in multiple national and international research projects. From the beginning, it was

designed to use standardized file formats, thus ensuring interoperability with other language processing systems. Internal functions are available as JSON REST web services, thus allowing for interaction with other systems. In the Representational State Transfer (REST) architectural style, data and functionality are considered resources and are accessed using Uniform Resource Identifiers (URIs).

This paper is organized as follows: section 2 presents related work in the field of LT platforms; section 3 presents the history of how the RELATE platform developed; section 4 introduces RELATE's architecture; section 5 describes the platform's components; section 6 presents several scenarios in which the platform might be further used; and we draw conclusions in section 7.

2. Related work

META-SHARE[1] (Federmann *et al.* 2012), CLARIN[2] and ELRC-Share[3] (European Language Resource Coordination Share) are publicly available European websites for research and development in the field of natural language processing, allowing access to language resources. Both ELRC-Share and META-SHARE offer advanced search facilities through which one can easily find various language tools and corpora (e.g. text corpora, annotated corpora, audio corpora) in any (European and other) language. Complex processing pipelines, like NLP-Cube (Boroș *et al.* 2018a) and TTL are able to perform tokenization, lemmatization, POS tagging, chunking and dependency parsing, but they require programming knowledge in order to integrate these functions into an application. A more recently created web service, TEPROLIN, integrates multiple tools — including NLP-Cube, TTL and solutions for named entity recognition (Păiș 2019) and biomedical named entity recognition — into an easy-to-use pipeline. However, this web service still requires users to have programming knowledge in order to engage in web service communication.

GATE (Cunningham *et al.* 2002) and TextFlows (Perovšek *et al.* 2016) aim to make the composition of the language processing chains more user-friendly. The

[1] http://www.meta-share.org.
[2] https://www.clarin.eu.
[3] https://elrc-share.eu/.

graphical interfaces thus allow text-processing widgets to be dragged and dropped into a graphical processing workflow. However, text processing output is not enhanced with specialized visualization tools that would allow access to the computational resources used for annotation. Moreover, they only focus on purely textual content, while multimodal content (such as a combination of text and audio) is not handled within the platforms.

CoBiLiRo (Cristea *et al.* 2020) is a storage platform for multimodal (text and audio) corpora. It was developed in the context of the RETEROM[4] project for storing Romanian speech corpora in a way that was suitable for the project's purposes. It allows a large number of metadata fields to be defined, but it does not enable any complex language processing tasks.

RELATE specifically serves to carry out automatic text processing, with annotations at multiple levels, as well as annotation visualization and expansion into the corresponding linguistic computational resources. The platform focuses on the interactive user experience, through web-based interfaces. Unlike other platforms, such as the one designed by Che *et al.* (2010), RELATE is not primarily focused on exposing APIs, even though API access to platform functionalities is available. Text-processing APIs are used internally and are accessible by platform components. Most of these APIs can also be invoked externally.

RELATE makes use of a common internal format, comprising multiple files (text, standoff metadata, CoNLL-U Plus[5] annotations). The processing workflow is built by adding individual tasks in the graphical user interface. Each task is able to both use and also produce data according to the common internal format. For this reason, no workflow editor, such as the one used in TextFlows (Perovšek *et al.* 2016), is currently available.

Coleman *et al.* (2020) describe a platform for integrating multiple Machine Translation (MT) models. In RELATE, we provide those MT capabilities that are related to the Romanian language (currently supporting Romanian-English and English-Romanian translation only). Furthermore, in RELATE, translated content

[4] https://racai.ro/p/reterom/.
[5] https://universaldependencies.org/ext-format.html.

can be used as input for further language processing tasks. Rebai *et al.* (2020) describe a platform that integrates a voice assistant for improving efficiency and productivity in business. In the case of RELATE, we only integrate existing Romanian (and English) Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems while still focusing on the textual components of a corpus. Thus, ASR output can be used as input to further language processing tasks.

Rehm *et al.* (2020b) acknowledge that a large number of AI platforms are currently being developed, both nationally—supported through state funding programmes—and internationally, supported by the European Union. The authors further recognize the enormous fragmentation of the European AI and LT landscapes and believe that modern platforms should facilitate exchange information, data and services, in order to enable interoperability. We agree with this assessment and, in the RELATE platform, we therefore sought to make use of standardized formats as well as decoupling functionalities into components, which could be invoked externally if needed. In addition, similar to the AI4EU[6] and European Language Grid (ELG)[7] platforms, RELATE is able to integrate Docker containers for language processing tools. Such integration is not required, however, and only a small number of the available components are therefore integrated as containers.

3. Evolution of the RELATE platform

The RELATE platform was developed in the context of a number of national and international research projects and evolved according to the needs of the activities for which it was ultimately used. This section describes its evolution throughout these projects, from the first implementation to the current version.

RELATE platform development (Păiș *et al.* 2019) started in the context of the national RETEROM project, which began in 2018. One of the project's main goals, linked to the sub-project TEPROLIN, was to develop and integrate state-of-the-art technologies for Romanian natural language processing, such as those devoted to tokenization, part-of-speech tagging, dependency parsing, phonetic annotations, and

[6] https://www.ai4eu.eu/.
[7] https://www.european-language-grid.eu/.

named entity recognition. The primary result was the TEPROLIN web service (Ion 2018) which allows invocation that uses a raw text document and produces different levels of annotation (based on specified parameters) encoded using a custom JSON format.[8]

The first implementation of the RELATE platform allowed for the management (upload, download, editing) of large corpora and parallel processing using the TEPROLIN service. For processing purposes, we implemented a task management engine, which allowed documents to be distributed across any number of TEPROLIN processes that begin on the same server or over the network. As a result, the processing speed can be increased when needed, because the number of processing nodes can be adjusted dynamically. Furthermore, the platform is able to convert the custom JSON encoding associated with TEPROLIN into a more standard CoNLL-U format[9] (also used by the Universal Dependencies project[10]) or its extension CoNLL-U Plus,[11] when appropriate.

To allow human users to manually explore annotation results, we developed graphical representations, such as visualization in different column-based and JSON formats, tree-based representations for dependency parsing, and the highlighting of recognized named entities. In order to improve the user experience, we also included different interface elements that allow the Representative Corpus of Contemporary Romanian Language (CoRoLa) to be queried (Tufiș *et al.* 2019b). RELATE includes links to the main query interface of CoRoLa's text component, which are accessible through the KorAP corpus analysis platform (Banski *et al.* 2012), and to the speech component, which allows users to search audio files (Boroș *et al.* 2018b) and listen to words being pronounced by speakers of Romanian. Different visualizations also make use of word embeddings (Bojanowski *et al.* 2017) computed on the CoRoLa corpus (Păiș and Tufiș 2018) to suggest words that appear in similar contexts.

We integrated the Romanian WordNet (Tufiș and Mititelu 2015) into the RELATE platform to allow words to be searched online. Furthermore, we exploited

[8] http://relate.racai.ro/?path=teprolin/doc_dev.
[9] https://universaldependencies.org/format.html.
[10] https://universaldependencies.org/.
[11] https://universaldependencies.org/ext-format.html.

the alignment between the Romanian and English (Miller 1995) wordnets in order to provide aligned queries based on *synset* identifiers (in a WordNet, synsets represent sets of synonyms that share a common meaning).

A textual translation component—which was developed by TILDE with the involvement of ICIA, in the context of the project entitled "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU"—was integrated into RELATE through the TILDE Machine Translation API.[12] This component allows users to translate documents directly within the platform. After a successful translation from English to Romanian, the resulting document can be analyzed using the platform's text analysis functionality.

As part of the "Multilingual Resources for CEF.AT in the legal domain" project (MARCELL),[13] a terminology annotation tool (Coman *et al.* 2019a) was developed, and we subsequently integrated it into the RELATE platform to facilitate the identification of terms available in the EuroVoc[14] and IATE[15] (Interactive Terminology for Europe) terminological databases. This annotation tool can be used following a lemmatization and part-of-speech tagging operation.

The ASR system (Avram *et al.* 2020b) that resulted from the national project ROBIN[16] and was initially developed for human-robot interaction (Ion *et al.* 2020; Tufiș *et al.* 2019a) was also integrated into the RELATE platform. This system allows text to be extracted from speech recorded directly in the platform or from uploaded audio files and then to be analyzed with the platform's processing tools. Several corrections on the recognized text (such as basic comma restoration and truecasing) can be performed before the actual analysis is carried out (Avram *et al.* 2020a).

We further exploited the availability of ASR and text translation features in order to provide speech-to-speech translation functionality for Romanian-English and English-Romanian. This brought together different platform modules and

[12] https://www.tilde.com/developers/machine-translation-api.
[13] https://marcell-project.eu/.
[14] https://eur-lex.europa.eu/browse/eurovoc.html.
[15] https://iate.europa.eu/home.
[16] http://aimas.cs.pub.ro/robin/en/.

required the integration of additional text-to-speech components. To synthesize the Romanian speech, we integrated two models into our pipeline: Romanian TTS (developed by Stan *et al.* (2011)) and RACAI SSLA (developed by Boroș *et al.* (2018b)), both of which are based on Hidden Markov Models (HMM) to compute the most probable sequence of spectrograms.

The "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project, launched in 2021, requires large-scale anonymization of the collected text corpus. This process was further integrated into RELATE, making use of the parallelization mechanism that was already available for different text-processing tasks.

One of our institute's research objectives, which aligns with the COST Action "Nexus Linguarum", is to create resources that are specific to Linguistic Linked Open Data specifications (Barbu Mititelu *et al.* 2020). Our involvement in the COST Action provided an opportunity to develop post-processing capabilities within RELATE, allowing RDF exports for different annotation levels.

The "European Language Equality" (ELE) project,[17] initiated in 2021, has as its primary goal to develop a sustainable evidence-based strategic research agenda and a roadmap that sets out actions, processes, tools, and actors that are integral to achieving—through the effective use of language technologies—digital equality among all of the languages (official or otherwise) used within the European Union. Additionally, the research agenda encompasses the assessment of the current state of language technologies and language equality within the EU. In turn, this led to an extension of RELATE, allowing a number of similar text-processing pipelines (Păiș 2020) to be executed on the same corpus in order to compare the results of each one (with or without the presence of a gold-standard annotated corpus).

4. Platform architecture

We adopted a modular approach to organize the RELATE platform. Each component was implemented independently of the others and provides a JSON descriptor that exposes the corresponding menu entries and functionalities at

[17] http://www.european-language-equality.eu.

different user access levels. Moreover, components are service oriented. They can consume web services (either SOAP or REST) and expose additional HTTP REST services. In turn, the exposed services can be consumed by the component's graphical user interface, through AJAX calls to the REST endpoints.

This service-oriented architecture allows for easy reuse of component-provided functionality in other parts of the platform or even from outside the platform. By consuming services exposed by processing components, RELATE can deploy these functions on different computing nodes, thus enabling high-speed parallel processing across interconnected servers. The processing parallelization was used for large corpora annotation in our institute, engaging multiple servers available in the same local area network (LAN). However, since a processing operation takes more time than network communication, it is possible to use remote hardware resources (across the internet) to further increase the processing capabilities available to the platform.

Figure 1 presents RELATE's general platform architecture, identifying its various components. The front-end part of the platform is responsible for user interaction. It provides visualizations of the different components. It is implemented using modern front-end technologies and languages such as HTML5, CSS, JavaScript. Communication with the platform's back end occurs through page loading or AJAX calls when possible. The use of AJAX enhances the user experience by reducing loading times. Furthermore, exposing different APIs in the form of HTTP REST services enables direct usage from outside the platform if needed.

The RELATE back end is responsible for communicating with different language-specific components. The back end itself is written in PHP, while the components can be written in any language, as long as a communication mechanism is available. The platform currently has components written in C++, Java and Python. The preferred means of communication is HTTP REST API calls. Nevertheless, the platform also integrated components by invoking new processes with command line arguments.

5. Primary platform components

This section will present the primary platform components (as indicated in *Figure 1*), addressing topics such as the platform's functions and implementation details. Only components that have undergone recent developments are included. Older components, such as the WordNet and CoRoLa query interfaces are not discussed in detail. Additionally, some of the more minor components will be discussed in Section 6.

5.1. Corpus annotation

TEPROLIN (Ion 2018) is a text pre-processing platform, implemented as a web service. It was developed as part of the ReTeRom project,[18] with the specific goal of making it easy to bring together diverse text pre-processing tools that are written in a host of programming languages. TEPROLIN is written in Python 3 and requires each participating text pre-processing application to implement the TEPROLIN API class.[19] Ensuring that the following conditions are met leads to best results in TEPROLIN:

- External resources should always be loaded at the same time as the object is created; this ensures that, with large resource files, the loading time does not interfere with the run time.
- Text pre-processing applications should be resident processes on the machine that is running TEPROLIN, and the I/O with TEPROLIN should be made with available IPC mechanisms, such as TCP sockets.
- If text pre-processing applications are designed as web services, no resource loading is necessary and the _runApp method is the only one that needs to be implemented.

TEPROLIN can be used either as a stand-alone Python 3 object or through the RELATE platform, in demo[20] or production mode (for annotating large corpora). Each processing operation (e.g. POS tagging, dependency parsing) can be configured individually. If desired, only a subset of operations can be requested to be performed. In this case, TEPROLIN will auto-configure to execute all the other required

[18] https://www.racai.ro/p/reterom/.
[19] https://github.com/racai-ai/TEPROLIN/blob/master/TeproApi.py.
[20] https://relate.racai.ro/index.php?path=teprolin/custom.

operations. For instance, if the user requires POS tagging, TEPROLIN will automatically conduct sentence splitting and tokenization beforehand.

Currently, TEPROLIN includes the following text pre-processing applications:

- NLP-Cube (Boroș *et al.* 2018a), UDPipe (Straka *et al.* 2016) and TTL (Ion 2007)
  - All three carry out sentence splitting, POS tagging and lemmatization functions.
  - NLP-Cube and UDPipe perform dependency parsing.
  - TTL performs noun phrase, verb phrase, prepositional phrase, adverbial and adjectival phrase non-recursive chunking.
- MLPLA (Boroș *et al.* 2018b)
  - This application performs word hyphenation, stressed accented syllable detection, and phonetic transcription.
- Various in-house tools
  - These were developed for Romanian text normalization, diacritic restoration, abbreviation and numeral expansion (i.e. automatically derive the full form of the abbreviation or the literal transcription of the numeral).
- BioNER (Mitrofan *et al.* 2018)
  - This application is based on a previous version of NLP-Cube adapted for NER processing.
- NER (Păiș 2019)
  - This application detects names of persons, organizations, locations and time expressions.

UDPipe[21] (Straka *et al.* 2016) is a trainable application that is used for the tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. Sentence splitting and tokenization are performed by checking each character in the input—using a single-layer bidirectional GRU network—to determine whether the character is the end of a token, the end of a sentence, or both. POS tagging and

[21] https://github.com/ufal/udpipe.

lemmatization use an averaged perceptron with morphosyntactic features (e.g. word affixes paired with current POS tags), while the dependency parser uses a simple feed-forward neural network to select the optimal transition indicated by the chosen transition-based dependency parser. Instead of TEPROLIN, UDPipe can be used as a stand-alone text pre-processing pipeline in RELATE, within its production mode (i.e. using multiple processing threads to annotate large corpora).

BioNER annotation is obtained through TEPROLIN, through requests for the biomedical-named-entity-recognition operation. It produces inside-outside-beginning annotations of disorders (DISO), anatomical parts (ANAT), medical procedures (PROC), drugs and other chemicals (CHEM). Mitrofan *et al.* (2018) describe the corpus that was used to train a previous version of NLP-Cube to recognize these specialized text spans.

Another type of NER annotation involves using a modified version of an older Stanford Named Entity Recognizer (Stanford NER) (Finkel *et al.* 2005) system, making use of CoRoLa pre-trained word embeddings (Păiș and Tufiș 2018), and computing representations of unknown words based on sub-word information (Bojanowski *et al.* 2017). The annotation is further enhanced through a rule-based method. Finally, results are attached to a tokenized document using Inside-Outside (IO) notation, which is then converted to an Inside-Outside-Beginning (IOB) format for compatibility with other annotations.

Besides TEPROLIN and UDPipe processing, pre-trained language models for other basic language resource kits (BLARK) are available for download from within the RELATE platform.[22] These models are trained on the same corpus, namely version 2.7 of the RoRefTrees treebank (RRT) (Barbu Mititelu, 2018), which is accessible through the Universal Dependencies[23] project. Models are offered for

[22] https://relate.racai.ro/index.php?path=pretrainedlm/list.
[23] https://universaldependencies.org/.

Stanza[24] (Qi *et al.* 2020), RNNTagger[25] (Schmid 2019), NLP-Cube[26] (Boroș *et al.* 2018a), UDPipe (Straka *et al.* 2016) and TreeTagger[27] (Schmid 1994).

5.2. Terminology annotation

The terminology annotation tool available in RELATE was initially created for the purposes of the international project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL).[28] This project produced a large comparable corpus (Váradi *et al.* 2020) of national legislation from the seven partners' countries, including Romania, enriched with EuroVoc and IATE terminology annotations. Furthermore, a classification of these legislative documents based on the top-level domains of the EuroVoc multilingual thesaurus was carried out.

For categorizing documents belonging to the Romanian-language sub-corpus (Tufiș *et al.* 2020), we developed a classifier based on the method proposed by Joulin *et al.* (2017), using the average of word embeddings over n-grams as features to a linear classifier. The actual implementation used the FastText[29] tool. We trained several models, using different word embeddings (CoRoLa, Wikipedia and CommonCrawl). We wanted to compare these models with the well-known JEX tool,[30] developed by the European Commission's Joint Research Centre (EC-JRC). JEX (Steinberger *et al.* 2012) makes use of the classification algorithm described in Pouliquen *et al.* (2006), based on a list of lemma frequencies from normalized text, and their weights, which are statistically related to each descriptor, mentioned in the paper as associates or as topic signatures. For Romanian, JEX was trained on over 25,000 documents from two parallel corpora: JRC-Acquis (Steinberger *et al.* 2006) and OPOCE (Publications Office of the European Union). By considering only the first 6 most probable descriptors, it obtained a precision of 45.55%, a recall of 50.43% and a F1 score of 47.84%.

[24] https://stanfordnlp.github.io/stanza/index.html.
[25] https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/.
[26] https://github.com/adobe/NLP-Cube.
[27] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
[28] https://marcell-project.eu/.
[29] https://fasttext.cc/.
[30] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer.

For training and evaluating our EuroVoc models, we used the same approach followed by JEX, employing the same datasets and cross-validation splits. Our best model was obtained through the use of CoRoLa word embeddings and achieved a precision of 50.93%, a recall of 56.41% and a F1 score of 53.53%, thus presenting an increase of over 5% with regard to F1 score as compared to JEX. We further evaluated the model on top-level domains and obtained an F1 score of 70.80%. The model is available for download and interrogation from RELATE.[31] In order to integrate this model into the platform, we employed a modified version of the FastText application, including an embedded web server, which allowed the model to be kept in-memory. This server is open-sourced.[32] Besides document classification, term annotation was realized through the use of a NER-based approach, as described by Coman *et al.* (2019b).

For the named entities belonging to the location class (LOC) of the LegalNERo corpus[33] (Păiș *et al.* 2021b), we carried out an enhancement that linked the named entities with the GeoNames database[34] by using the feature identifiers associated with each GeoNames feature. GeoNames is a geographical database that contains over 25 million geographical names, covering all countries, and consists of over 11 million unique placenames, 4.8 million populated places and 13 million alternate names. GeoNames is integrating geographical data such as elevation, population, names of places in various languages, and other data from various sources. To link the named entities with the GeoNames database, we created a new column in the conllup corpus files structure. This GeoNames column was placed last in the file and is, by default, filled with the "_" character. This column contains a value only when the named entity tag is of LOC type and when a perfect match is found within the GeoNames database. The actual match is done through a lookup algorithm written in the Python language. In this algorithm, the GeoNames database extract for the Romanian language is loaded into a *key:value* list, where the key is the standard location name and the value is the GeoNames location code (e.g. the code for the city

[31] https://relate.racai.ro/index.php?path=eurovoc/classify.
[32] https://github.com/racai-ai/ServerFastText.
[33] https://doi.org/10.5281/zenodo.4772094.
[34] https://www.geonames.org/.

of Suceava is 486885). In many cases, a given location can have several variations in the naming (whether on account of different spelling conventions or due to differences in how the places are known across and within different communities). GeoNames addresses this phenomenon by listing only one standard name (usually the location's official name) and including several secondary names alongside the GeoNames entry for that location. For each variation of this kind, the application has therefore created a different entry in the lookup dictionary, all of them pointing to the same location code value. The conllup file format is token centric, while certain locations have a multi-word format. Therefore, the first step in the lookup was actually to build the expressions based on the tokens (by finding consecutive LOC tokens). The actual lookup was performed for each named entity (a multi-token list). When a perfect match was found (partial sub-expression matches were ignored), all of the tokens involved in the match were flagged as a GeoNames entity. The corresponding location code was finally entered into the GeoNames column.

5.3 Automatic speech recognition

The RELATE platform contains two Romanian ASR models, which can be accessed using the graphical interface: RobinASR and RobinASRDev. The two models are based on the DeepSpeech2 architecture (Amodei *et al.* 2016) and, as compared with the original model, have a reduced number of parameters, which allows the smaller size of their training corpus to be accommodated. Each model is wrapped as a web service and deployed on a server with a GPU, allowing a low response time to the request on the front end to be achieved. Although the two variants are similar in architecture, however, they serve different purposes.

The former variant is a general ASR that was trained on a corpus containing 230 hours of transcribed audio and that obtained a 9.91 word-error rate (WER) when combined with a language model (Heafield 2011) used for textual correction. The latter variant is a specialized ASR. It fine-tunes the general model on the technical domain using the ROBIN Technical Acquisition Speech Corpus (ROBINTASC) corpus (Păiș *et al.* 2021a) and improving the model's performance by 16.4% on computer sales conversations. This variant also uses a language model that was

trained on the general corpus together with the train ROBINTASC transcriptions, oversampled 10 times to increase n-grams frequencies.

5.4 Text and speech translation

The speech-to-speech translation (S2ST) system (Avram *et al.* 2021b) consists of four cascaded modules: (1) ASR, (2) Textual Correction (TC), (3) Machine Translation (MT) and (4) TTS. Each module contains one or several models that can be configured in a full Romanian-English or English-Romanian inference, as depicted in *Figure 2*.

The ASR module contains two models for Romanian—namely those presented in Subsection 5.3—and two models for English. The first variant for English (En DeepSpeech2) is also based on the DeepSpeech2 architecture and was trained on the LibriSpeech dataset (Panayotov *et al.* 2015), obtaining a 10.46 WER. The second variant for English (Mozilla DeepSpeech) is the 0.9.3 version of DeepSpeech (Hannun *et al.* 2014) offered by Mozilla Speech-To-Text engine,[35] and it obtained a 7.06% WER on LibriSpeech.

At this time, we offer only a single model in the TC module for Romanian: Robin Correction. Its role is (1) to truecase the transcribed speech using a list of predefined named entities, (2) to replace unknown words with words from an existing vocabulary and (3) to restore hyphens using uni-gram and bi-gram statistics.

The Romanian-English (Presidency Ro-En) and English-Romanian (Presidency En-Ro) models used by the MT module are those that were developed in the context of the "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU" project, described in section 3.

The TTS module contains a speech synthesizer model for English and two for Romanian. The English version—Mozilla En TTS—is based on Tacotron2 with Dynamic Convolution Attention (Battenberg *et al.* 2020) offered by Mozilla Text-to-Speech.[36] Their evaluation outlined a median opinion score (MOS) of 4.31 ± 0.06 in a 95% confidence interval. The Romanian variants—Romanian TTS and RACAI

[35] https://github.com/mozilla/DeepSpeech/releases/tag/v0.9.3.
[36] https://github.com/mozilla/TTS.

SSLA—use HMM on their architecture, with the former variant being slower, but producing a higher quality speech and the latter being faster but with a lower quality of the synthesis.

6. Usage scenarios

Due to its complex nature and its large number of integrated components, the RELATE platform has been used for a variety of purposes. However, the most common ones are described in this section. Moreover, the description of the usage scenarios provides insights into possible interactions between the platform's components.

6.1 Large corpus processing

Large corpora can be uploaded to RELATE as archived zip files. Once they are uploaded, the platform will automatically schedule a task for extracting the archive. Inside the archive, the platform accepts raw text files (with .txt extensions) and standoff metadata. Metadata are usually specific to different research projects and can be used by different processing components to embed them into annotated documents.

Once the documents are available in the platform, different annotation tasks can be scheduled and executed. The tasks interface is presented in *Figure 4*. Furthermore, statistics on both raw text and annotated documents can be computed. The resulting annotated corpus can be archived and downloaded as another zip file. Statistics can be either visualized within the platform or downloaded as csv files. The entire flow is depicted in *Figure 3*.

6.2. Creation of human-annotated gold-standard text corpora

Creating a gold-standard corpus involves human annotators identifying spans of text with certain properties. For the creation of the LegalNERo[37] corpus (a gold-standard corpus for named entity recognition in the Romanian legal domain), we

[37] https://doi.org/10.5281/zenodo.4772094.

used the services of five human annotators who worked with RELATE and annotated spans of text with the corresponding named entities: legal reference, person, location, organization and time (Păiș *et al.* 2021b).

For the purpose of manual annotation of text corpora, RELATE integrates the BRAT[38] annotation tool (Stenetorp *et al.* 2012). This allows the user to view one document at a time inside the BRAT component, use their mouse to select the desired text span and then associate a corresponding annotation type with it. The text spans with an associated start and end index as well as the annotation type are saved as standoff metadata files.

Since different corpus processing applications may require token-based annotations instead of span-based annotations, RELATE offers dedicated tasks for parallel conversion of the metadata to token-based annotations. This process must be executed after an initial tokenization of the data, possibly including other automatic annotations such as part-of-speech tagging and dependency parsing. The final format for the data will be CoNLL-U Plus files with the gold-standard annotations stored in a dedicated column, named "RELATE:NE."

Once the corpus has been annotated (with or without conversion to the token-based format), it becomes possible to execute additional tasks within RELATE. A dedicated task allows extraction of gazetteer resources, allowing the produced file to be downloaded and used in other applications that require such resources. In addition, the statistics component can be used to extract information about the annotated corpus. The processing flow is depicted in *Figure 5*, while an example of the manual annotation component used with BioNER entities is presented in *Figure 6*.

6.3. Creation of read-speech corpora

Read speech corpora represent invaluable resources for the creation of ASR and TTS systems. These systems can then be used as the building blocks for more advanced processing, like human-machine interaction and speech-to-speech translation. In the context of the ROBIN project, we were concerned with the

[38] https://brat.nlplab.org/index.html.

performance of a low-latency ASR system used for human-robot interaction, in a micro-world scenario. For this purpose, we created the ROBIN Technical Acquisition Speech Corpus (RTASC)[39] (Păiș *et al.* 2021a), using the audio recording features of the RELATE platform.

Even though RELATE was initially constructed to process text corpora, the recent addition of speech processing features has allowed it to handle bimodal (text and audio) corpora. For recording purposes, a dedicated component was built. It makes use of JavaScript and runs on the user's browser. It displays each sentence from the text component of the corpus and allows the user to record the associated speech. To improve the user experience, RELATE displays the sentence in a larger font and provides additional information, such as the current sentence number and the total number of sentences. The interface also provides a playback function and, if necessary, allows the user to delete a certain recording in order to record it again. The recording interface is shown in *Figure 8*.

Since the text component still plays an important role in bimodal corpora, all the text-processing capabilities can be used to enhance the final corpus. In this case, the annotation pipelines and the statistics generation features are available. The final corpus, comprised of raw text, annotated text and audio files, can be downloaded in an archived format from the platform. The entire processing chain is presented in *Figure 7*.

7. Conclusions

In this article, we described the RELATE platform's development and architecture, and we considered several of the most common usage scenarios. Currently, the platform provides different levels of processing for the Romanian language. Even though RELATE was initially developed for handling large text corpora, recent work has integrated speech capabilities, including speech-to-speech translation for Romanian-English and English-Romanian language pairs.

[39] https://doi.org/10.5281/zenodo.4626539.

The platform is designed to be highly customizable and easily extensible thanks to the creation of new components. The use of standardized file formats ensures interoperability with other systems. Also, the extensive use of REST APIs allows interoperability with external applications. Finally, the platform is open-sourced and available free of charge on GitHub.[40] The current running version of RELATE[41] spans its processing capabilities across four servers hosted at ICIA. The servers contain both classical CPUs and a GPU resource, allowing different types of language processing tools to be used and guaranteeing high performance.

In the future, we plan to continue integrating new language processing tools into the RELATE platform, as they become available, while aiming to boost interoperability and interactions leading to a productive LT ecosystem (Rehm *et al.* 2020a), both nationally and internationally. One possible extension in this direction is the replacement of the current EuroVoc annotator with the Romanian model of PyEuroVoc (Avram *et al.* 2021a), a new legal document classification tool that works with 22 languages that is based on a more modern architecture—Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.* 2019)—and that significantly improved the results of JEX by an average of 28% F1 score for all languages and 33.8% for Romanian.

[40] https://github.com/racai-ai/RELATE.
[41] https://relate.racai.ro.

Annex 1: Figures

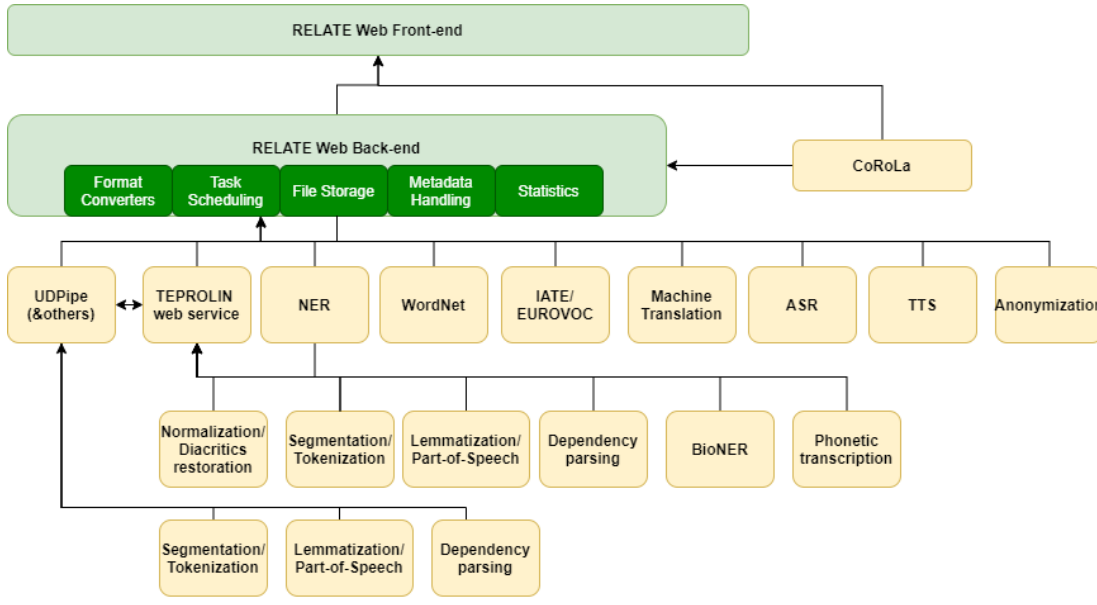


Figure 1: RELATE platform architecture

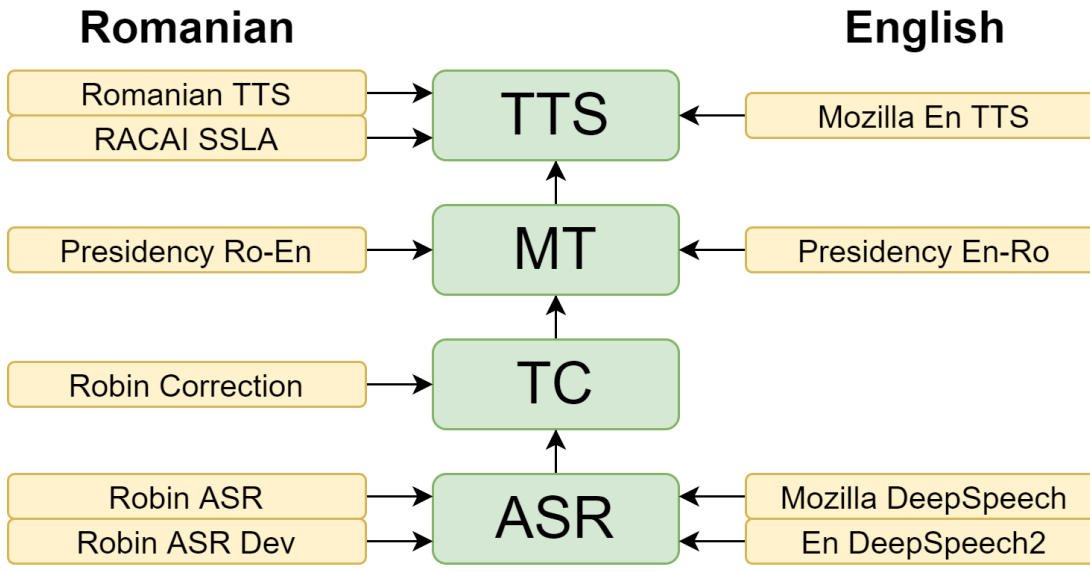


Figure 2: The S2ST cascaded architecture with the four modules and their models (Romanian in the left, English in the right).

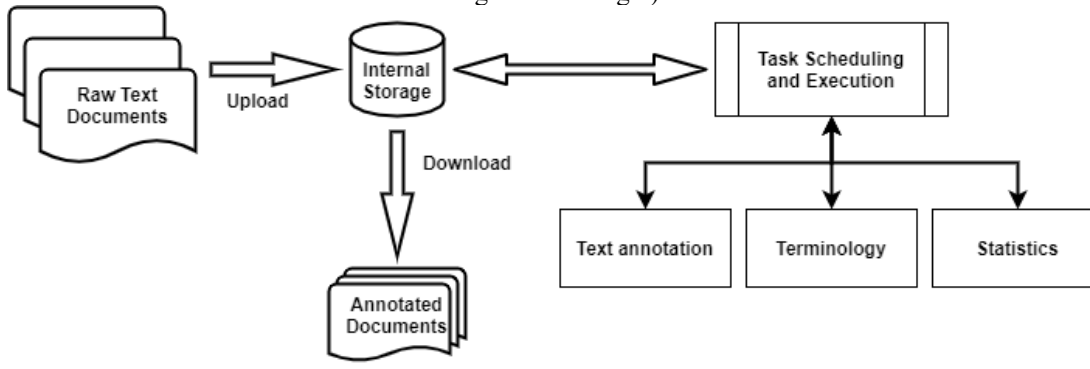


Figure 3: Corpus processing flow

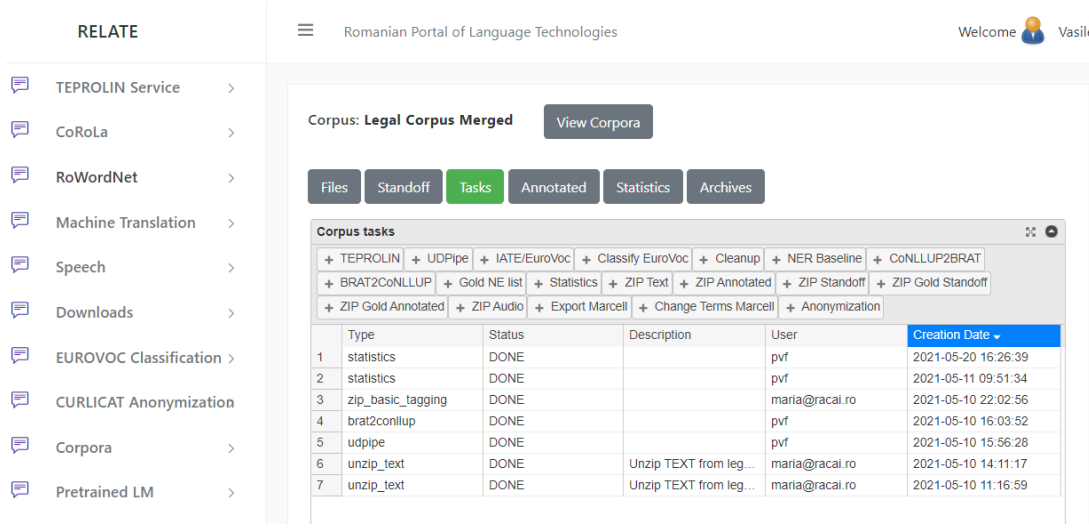


Figure 4: Tasks interface.

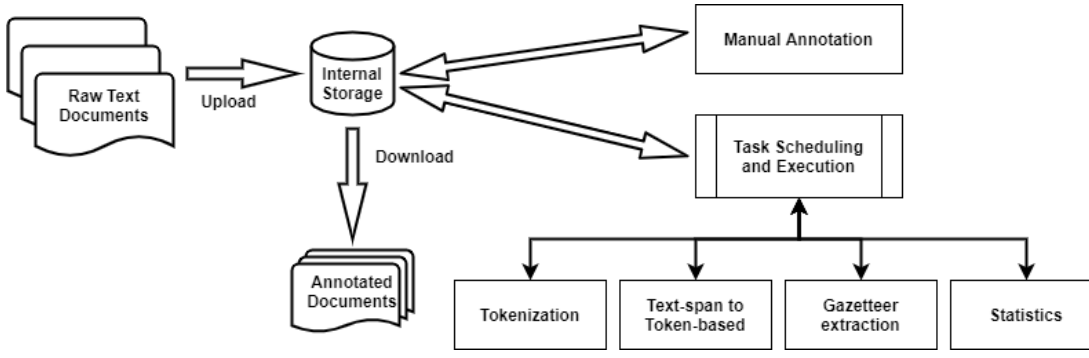


Figure 5: Manual corpus annotation processing flow

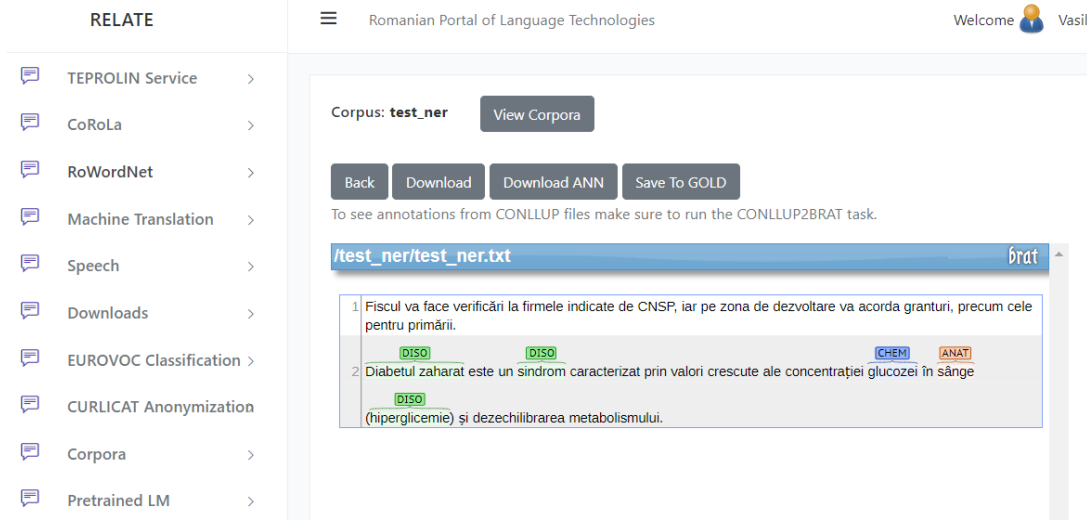


Figure 6: Manual corpus annotation GUI

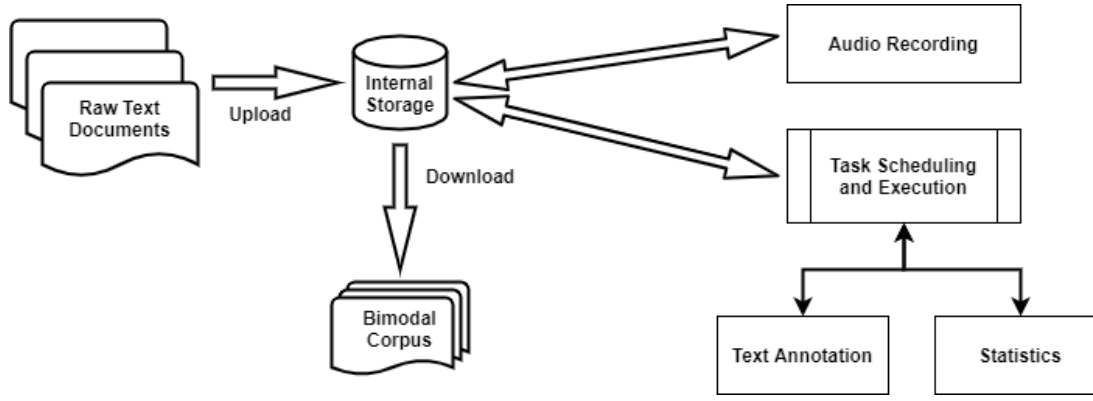


Figure 7: Creation of a bimodal corpus

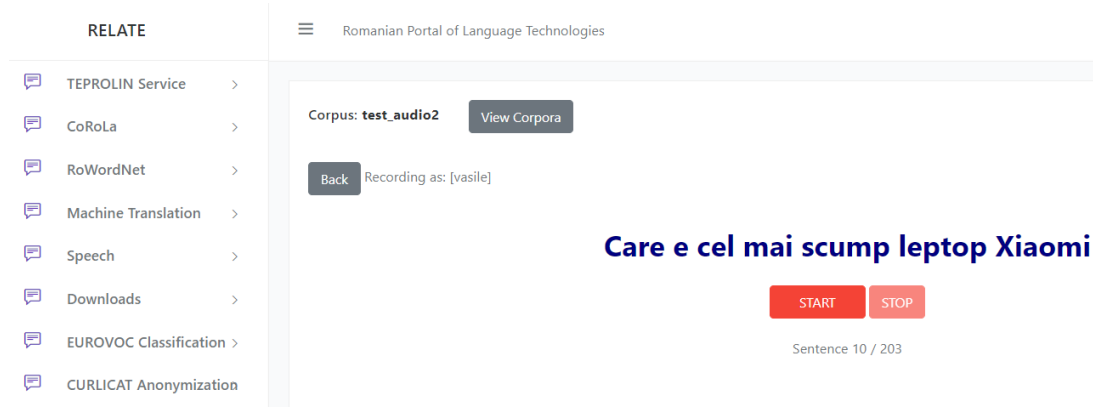


Figure 8: The interface used for recording read speech

References

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, *et al*. 2016. "Deep speech 2: End-to-end speech recognition in English and Mandarin." In *International Conference on Machine Learning*, 173–182. PMLR.

Avram, Andrei-Marius, Vasile Păiș, și Dan Tufiș. 2020a. "Romanian speech recognition experiments from the ROBIN project." In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 103–114.

Avram, Andrei-Marius, Vasile Păiș, and Dan Tufiș. 2020b. "Towards a Romanian end-to-end automatic speech recognition based on DeepSpeech2." *Proceedings of the Romanian Academy Series A* 21:395–402.

Avram, Andrei-Marius, Vasile Păiș, and Dan Tufiș. 2021a. "PyEuroVoc: A tool for multilingual legal document classification with EuroVoc descriptors." arXiv preprint arXiv:2108.01139.

Avram, Andrei-Marius, Vasile Păiș, and Dan Tufiș. 2021b. "A modular approach for Romanian-English speech translation." In *International Conference on Applications of Natural Language to Information Systems*, 57–63. Springer.

Banski, Piotr, Peter M. Fischer, Elena Frick, Erik Ketzan, Marc Kupietz, Carsten Schnober, Oliver Schonefeld, and Andreas Witt. 2012. "The new IDS corpus analysis platform: Challenges and prospects." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2905–2911. Istanbul, Turkey: European Language Resources Association (ELRA).

Battenberg, Eric, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. 2020. "Location-relative attention mechanisms for robust long-form speech synthesis." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6194–6198. IEEE.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5:135–146.

Boroș, Tiberiu, Ștefan Daniel Dumitrescu, and Ruxandra Burtică. 2018a. "NLP-cube: End-to-end raw text processing with neural networks." In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 171–179. Brussels, Belgium: Association for Computational Linguistics.

Boroș, Tiberiu, Ștefan Daniel Dumitrescu, and Vasile Păiș. 2018b. "Tools and resources for Romanian text-to-speech and speech-to-text applications." In *Proceedings of the International Conference on Human-Computer Interaction (RoCHI)*, 46–53.

Che, Wanxiang, Zhenghua Li, and Ting Liu. 2010. "LTP: A Chinese language technology platform." In *Coling 2010: Demonstrations*, 13–16. Beijing, China: Coling 2010 Organizing Committee.

Coleman, Susie, Andrew Secker, Rachel Bawden, Barry Haddow, and Alexandra Birch. 2020. "Architecture of a scalable, secure and resilient translation platform for multilingual news media." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 16–21. Marseille, France: European Language Resources Association.

Coman, Andrei, Maria Mitrofan, and Dan Tufiș. 2019a. "Automatic identification and classification of legal terms in Romanian law texts." In *The 14th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 39–49.

Coman, Andrei, Maria Mitrofan, and Dan Tufiș. 2019b. "Automatic identification and classification of legal terms in Romanian law texts." In *International Conference on Linguistic Resources and Tools for Natural Language Processing*.

Cristea, Dan, Ionuț Pistol, Șerban Boghiu, Anca-Diana Bibiri, Daniela Gîfu, Andrei Scutelnicu, Mihaela Onofrei, Diana Trandabăț, and George Bugeag. 2020. "CoBiLiRo:

A research platform for bimodal corpora." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 22–27. Marseille, France: European Language Resources Association.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. "GATE: An architecture for development of robust HLT applications." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 168–175. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Federmann, Christian, Ioanna Giannopoulou, Christian Girardi, Olivier Hamon, Dimitris Mavroeidis, Salvatore Minutoli, and Marc Schroder. 2012. "META-SHARE v2: An open network of repositories for language resources including data and tools." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3300–3303. Istanbul, Turkey: European Language Resources Association (ELRA).

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating non-local information into information extraction systems by Gibbs sampling." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 363–370. Ann Arbor, Michigan: Association for Computational Linguistics.

Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, *et al*. 2014. "Deep speech: Scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567.

Heafield, Kenneth. 2011. "KenLM: Faster and smaller language model queries." In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 187–197.

Ion, Radu. 2007. "Word sense disambiguation methods applied to English and Romanian." PhD thesis, Romanian Academy in Romanian.

Ion, Radu. 2018. "TEPROLIN: An extensible, online text preprocessing platform for Romanian." In *The 13th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 69–76.

Ion, Radu, Valentin Gabriel Badea, George Cioroiu, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, and Dan Tufiș. 2020. "A dialog manager for micro-worlds." *Studies in Informatics and Control* 29(4):411–420.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. "Bag of tricks for efficient text classification." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Valencia, Spain: Association for Computational Linguistics.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Commun. ACM* 38 (11): 39–41.

Mitrofan, Maria, Verginica Barbu Mititelu, and Grigorina Mitrofan. 2018. "Towards the Construction of a Gold Standard Biomedical Corpus for the Romanian Language." *Data* 3 (4): 12.

Păiș, Vasile. 2019. "Contributions to Semantic Processing of Texts; Identification of Entities and Relations Between Textual Units; Case Study on Romanian Language." PhD thesis, Romanian Academy.

Păiș, Vasile. 2020. "Multiple Annotation Pipelines Inside the RELATE Platform." In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 65–75.

Păiș, Vasile, Radu Ion, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, and Andrei-Marius Avram. 2021a. "ROBIN Technical Acquisition Speech Corpus."

Păiș, Vasile, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. 2021b. "Romanian Named Entity Recognition in the Legal Domain (LegalNERo)."

Păiș, Vasile, Dan Tufiș. 2018. "Computing Distributed Representations of Words Using the CoRoLa Corpus." In *Proceedings of the Romanian Academy, Series A* 19 (2): 403–409.

Păiș, Vasile, Dan Tufiș, Radu Ion. 2019. "Integration of Romanian NLP Tools into the RELATE Platform." In *The 14th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 181–192.

Păiș, Vasile, Dan Tufiș, Radu Ion. 2020. "A Processing Platform Relating Data and Tools for Romanian Language." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 81–88. Marseille, France: European Language Resources Association.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. IEEE.

Perovšek, Matic, Janez Kranjc, Tomaž Erjavec, Bojan Cestnik, and Nada Lavrač. 2016. "TextFlows: A Visual Programming Platform for Text Mining and Natural Language Processing." *Science of Computer Programming* 121: 128–152. Special Issue on Knowledge-based Software Engineering.

Pouliquen, B., R. Steinberger, and C. Ignat. 2006. "Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus." *ArXiv*, abs/cs/0609059.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. Online: Association for Computational Linguistics.

Rebai, Ilyes, Sami Benhamiche, Kate Thompson, Zied Sellami, Damien Laine, and Jean-Pierre Lorre. 2020. "LinTO Platform: A Smart Open Voice Assistant for Business Environments." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 89–95. Marseille, France: European Language Resources Association.

Rehm, Georg, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs, eds. 2020a. *Proceedings of the 1st International Workshop on Language Technology Platforms*. Marseille, France: European Language Resources Association.

Rehm, Georg, Dimitris Galanis, Penny Labropoulou, Stelios Piperidis, Martin Wels, Ricardo Usbeck, Joachim Kohler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julian Moreno-Schneider, Florian Kintzel, Elena Montiel, Victor Rodriguez Doncel, John Philip McCrae, David Laqua, Irina Patricia Theile, et al. 2020b. "Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 96–107. Marseille, France: European Language Resources Association.

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of International Conference on New Methods in Language Processing*.

Schmid, Helmut. 2019. "Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts." In *New York, NY, USA*. Association for Computing Machinery, 133–137.

Stan, Adriana, Junichi Yamagishi, Simon King, and Matthew Aylett. 2011. "The Romanian Speech Synthesis (RSS) Corpus: Building a High Quality HMM-Based Speech Synthesis System Using a High Sampling Rate." *Speech Communication* 53 (3): 442–450.

Steinberger, Ralf, Mohamed Ebrahim, and Marco Turchi. 2012. "JRC EuroVoc Indexer JEX - A Freely Available Multi-label Categorisation Tool." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA), 798–805.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiș, and Dániel Varga. 2006. "The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy: European Language Resources Association (ELRA).

Stenetorp, Pontus, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Junichi Tsujii. 2012. "BRAT: A Web-based Tool for NLP-assisted Text Annotation." In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France: Association for Computational Linguistics, 102–107.

Straka, Milan, Jan Hajič, and Jana Straková. 2016. "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing." In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia: European Language Resources Association.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Radu Ion, and George Cioroiu. 2019a. "Making Pepper Understand and Respond in Romanian." In *22nd International Conference on Control Systems and Computer Science (CSCS)*, 682–688.

Tufiș, Dan, and Verginica Barbu Mititelu. 2015. "The Lexical Ontology for Romanian." In *Language Production, Cognition, and the Lexicon*, edited by Núria Gala, Reinhard Rapp, and Gemma Bel-Enguix, 491–504. Berlin and New York: Springer International Publishing.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei. 2019b. "Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus of Contemporary Romanian." *Revue Roumaine de Linguistique* 64 (3): 227–240.

Tufiș, Dan, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. 2020. "Collection and Annotation of the Romanian Legal Corpus." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2773–2777. Marseille, France: European Language Resources Association.

Váradi, Tamás, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. "The MARCELL Legislative Corpus." In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 3761–3768.

Tools and Resources for Digitizing Historical Romanian Documents

Victoria BOBICEV, Tudor BUMBU
Ludmila MALAHOV, Alexandru COLESNICOV
Svetlana COJOCARU, Liudmila BURTSEVA
Cătălina MĂRĂNDUC

1. Introduction

Digitizing and providing online access to historical literary and cultural treasures are top priorities in the Digital Agenda for Europe.[1] The Romanian language, considered a less commonly used language, is increasingly benefiting from the development of digital editions, specialized corpora, archives, and dedicated platforms for preserving and analyzing cultural heritage. Cultural policy initiatives in Romanian-speaking regions focus on the study, digitization, and preservation of their heritage, including Romanian texts in Cyrillic and mixed scripts/ alphabets. The digitization process entails overcoming challenges related to recognizing, correcting, and transliterating printed Romanian texts in Cyrillic and mixed scripts (Boian *et al.* 2014, Ciubotaru *et al.* 2015). A complex technology has been elaborated in order to support the digitization of historical Romanian heritage printed in the Cyrillic alphabet in the 17th–20th centuries (Colesnicov *et al.* 2021).

The endeavor to address the linguistic heritage of Romanian history involves tackling several specific challenges, including: (1) dealing with a multitude of language evolution periods; (2) coping with the scarcity of widely available resources; (3) managing the diverse array of alphabets used in historical printings, including mixed Cyrillic–Latin "transition alphabets" (Cazimir 2006); (4) overcoming the absence of reliable tools for accurately recognizing Cyrillic letters from various historical eras; (5) addressing the shortage of lexicons suitable for the

[1] http://www.digitisation.eu/community/map-of-the-digitisation-landscape/

time periods of these resources. To surmount these obstacles, we have developed a comprehensive platform that seamlessly integrates a suite of software components that enable image processing, text recognition, and transliteration into contemporary Latin script. This platform has been finely tuned to handle text recognition and transliteration across different historical epochs and to accommodate variations in alphabets used in Romanian-language printings in both Romania and the Republic of Moldova. Follow-up efforts have been concentrated on annotating historical and other non-standard texts. Carefully checked morphological, syntactic, and semantic annotation of the collection of historical documents, regional folklore, and other non-standard text materials from Moldova and Romania resulted in the creation of the RoDia Dependency Treebank. In order for this valuable resource to be more widely accessible, its annotation has been transformed and a Non-Standard Romanian Universal Dependency corpus has been added to the collection of the Universal Dependency (UD) project.

The description of our work is organized into three main parts. The first part (*Exploring the Digitization of Old Romanian Printing*) introduces the collections of scanned historical Romanian documents and the challenges associated with digitizing them. The second part (*The Digitization Platform*) provides a detailed description of the new digitization platform, its functionality and the resources it uses. The third part (*Annotation and Conversion in UD*) covers the corpus annotation and is divided into three subsections: 4.1 concentrates on morphological tagging; 4.2 describes syntactic annotation; 4.3 gives an account of the transformation of the annotation in Universal Dependency format; 4.4 introduces the semantic annotation principles; the automated parsing experiments and their results are discussed in its subsection; 4.5. Finally, the conclusion completes the paper by listing the main achievements of the described work.

2. Exploring the Digitization of Old Romanian Printing

We work with old Romanian books written in the Cyrillic script. This is a challenge faced by both researchers of Romania and scholars of the Republic of Moldova. Since these countries previously constituted a single state, historical

documents (written in old Romanian Cyrillic) are common to them both (Cristea *et al.* 2018). The starting point for our work is the scanned text, *i.e.*, the text presented in the form of page images. These text images were sourced from two electronic libraries of texts: the Bucharest Digital Library[2] and the National Library of the Republic of Moldova.[3] The National Library holds a collection of approximately 21,000 old and rare books. An important source of scanned collections is its National Digital Library, known as *Moldavica,*[4] whose collections of old books printed in Slavonic characters consists of 1,194 scanned documents, and the collection of national magazines from 1867 to 1945 contains 4,276 documents. The National Digital Library of Romania[5] includes 724 old Romanian periodicals and 1228 old Romanian book and bibliophile resources in Romanian, Greek and Slavonic, among other languages (*Figure 1*).

17th century	*Noul Testament* [New Testament], 1648
18th century	*Fiziognomie* [Physiognomy], 1785 *Ducere către aritmetica* [Leading to Arithmetic], 1785 *De obste Gheografie* [General Geography], 1795 *Așezamânt* [Settlement], 1786
19th century	*Epistolariu* [Epistolary], 1841 *Gramatica românească* [Romanian grammar], 1835 *Legiuirea Caragea* [Caragea Law], 1818
20th century	*Folclor din părțile codrilor* [Folklore from the Parts of the Forests], 1973

Table 1: Scanned, recognized, and transliterated books.

Table 1 lists some of the linguistic sources with which we worked. The sources are of different historical periods starting with the oldest ones printed in the very first printing houses situated in Moldova in the 17th–18th centuries. Romanian Cyrillic fonts, especially from the selected epoch, are much less variable than Latin fonts. The usage of the Cyrillic script is connected with the Slavonic liturgical language of

[2] http://www.bibnat.ro/
[3] http://www.bnrm.md/
[4] moldavica.bnrm.md/
[5] http://digitool.bibnat.ro/R?RN=171659921

the Orthodox Church. *Figure 2* presents three samples of scanned books from different periods. The document quality can vary from perfect to almost unacceptable. The general problem is caused by the variations in writing and by language changes. The main objective is to make these documents accessible to the general public, by presenting them in a way that uses contemporary orthography.

The process of converting paper-based historical documents into searchable electronic files involves two obstacles that have not been fully overcome until now. The first obstacle is that historical typography is inconsistent and idiosyncratic. Due to their age, pages in old documents feature speckles, distortions, and stains. Also, historical fonts vary considerably, even within a given book, and they are more difficult to read than their contemporary counterparts. Furthermore, modern-day linguistic tools and resources often cannot detect the peculiarities of historical language variations. As a result—and especially since each text yields its own mix of features and problems—the quality of the text recognition varies dramatically. The second obstacle is the extent to which language, including spelling, has shifted since the historical documents were originally created. Texts in original historical orthography are unevenly supported by OCR software (for example, 18th-century Romanian Cyrillic writing).

For processing these historical documents, we propose a new digitization pipeline which enables recognition of historical Romanian heritage documents that were printed with Cyrillic letters in the 17th–20th centuries. This is based on the commercial OCR tool ABBYY Finereader (AFR)[6]. In comparison with open-source free OCRopus[7] and Tesseract,[8] AFR has numerous advantages. It offers users a convenient graphical shell from which the document's entire recognition cycle is available, starting with scanning. It provides complex image correction (for example, the correction of trapezoid distortion), page segmentation with automatic detection of the segment type (text, table, or picture), an analysis of the table structure, dehyphenation, manual editing after recognition. The user can choose to input a new language, specifying its alphabet and downloading an additional user dictionary. In

[6] https://pdf.abbyy.com/
[7] https://github.com/ocropus-archive/DUP-ocropy
[8] https://github.com/tesseract-ocr/tesseract

other cases, AFR can be trained on the challenging document, and the accumulated character patterns (OCR models) can be downloaded and uploaded for use in subsequent projects. OCRopus and Tesseract do not offer a single product that provides both a full recognition cycle and a graphical shell.

This technology supports virtual keyboards, fonts, a transliteration tool, and other possibilities. We have created additional Python programs such as a tool to select the historical period and the geographical region, where the text was printed, for instance Iași, Bucharest, Târgovişte, Bălgrad, Uniev, Sas Sebeş, Snagov, Buzău, etc. (Bumbu 2021, Cojocaru *et al.* 2017). In particular, for Bucharest, the system is trained in recognizing various fonts such as the Royal Typography's or the Bucharest Metropolitan Chair's.

3. The Digitization Platform

We have developed a Digitization Platform (DP) for digitizing old documents written in Romanian Cyrillic. DP is a web application that features an interactive graphical interface and a set of APIs designed and managed through Django Rest Framework.[9] This application offers seven consecutive steps to enable recognition of Romanian Cyrillic documents from the 17th–20th centuries, transliteration of the texts into Latin alphabet, editing of recognized/transliterated texts, and downloading or publication of the results (Bumbu 2023).

The platform includes:

- image processing engines, such as ScanTailor,[10] AFR, and OpenCV[11];
- OCR models for Cyrillic characters, trained on datasets gathered from documents printed in the 17th–20th centuries;
- a transliteration tool that converts from old Romanian Cyrillic into contemporary Latin script;

[9] https://www.django-rest-framework.org/
[10] https://scantailor.org/
[11] https://opencv.org/

- virtual keyboards specific to the alphabets used in previously mentioned periods, such as the Romanian Cyrillic alphabet, the transition alphabet, and the Soviet Cyrillic alphabet.

The DP architecture (*Figure 3*) consists of a set of modules that are organized into four functional groups. The first functional group (G1) deals with image processing. This group is equipped with two basic preprocessing modules—image binarization and resolution correction—and can be expanded through the integration of third-party applications like Scan Tailor, AFR, OpenCV, and GIMP. The second functional group (G2) consists of modules that handle the optical recognition of documents. Among the options we can mention the selection of the OCR model depending on the historical period, the selection of a word dictionary used during recognition, activation of the editing module for the recognized text or the use of the OCR exception dictionary. The OCR engine is based on AFR and has models trained with datasets collected from documents printed in the 17th–20th centuries. Users can also add new OCR models. These models use the FineReader XML format, which contains the OCR model configurations, training datasets, required alphabets, and word dictionaries. The third functional group (G3) contains components related to transliterating the recognized text. Primary modules in G3 ensure spelling updates (upon request), the use of dictionaries of exceptions for transliteration, editing of the resulting transliterated text with a spellchecker, and automatic text correction using an artificial intelligence system. Secondary modules in this group look after such tasks as replacing apostrophes with hyphens and deleting hyphens that had been used to break words across line.

The fourth functional group contains modules devoted to managing and publishing digitized documents. These tasks involve saving recognized/transliterated texts in various formats; downloading processed images; uploading original documents and processed images to the cloud; saving the state of the digitized object in the database; and expanding word dictionaries. The digitization process on the platform includes seven steps, as outlined below.

Step1: File upload

One or more files can be processed in a single digitization cycle. The following file types are supported: png, jpeg, tiff. The total size of all uploaded files must not exceed 700MB, and the size of each file is limited to 100MB. Files can be selected from the user's computer. When two or more files are selected, all of them will be processed with the same processing options, and the user is responsible for ensuring that the uploaded files are from the same period, use the same alphabet, and require the same image preprocessing options. If the user's document sets reflect multiple periods or require different image preprocessing options, then the sets should be processed in different digitization cycles (*Figure 4*).

Step 2: Preprocessing of uploaded images

In this step, the users can choose the preprocessing engine and preprocessing options needed for uploaded pages. Scan Tailor is the first preprocessing engine displayed in the list of options in step 2 and offers the widest range of options of all three engines proposed. Unlike with the other engines, two ways of using the Scan Tailor engine for image preprocessing are proposed (*Figure 5*).

Moving forward, the spotlight shifts to the preprocessing engine integrated into AFR. Currently, the testing phase is limited to exploring the fundamental image preprocessing functions, encompassing tasks such as rectifying image resolution discrepancies, identifying the optimal resolution for a given image, correcting automatic page orientation, transforming images to black and white, mitigating ISO noise,[12] and aligning text lines. The OpenCV engine, another formidable contender, affords users the flexibility to manually set the resolution of the pre-processed image. Additionally, it stands out by amalgamating filters within a single option, streamlining processes to clean the image of speckles and reduce ISO noise simultaneously.

[12] ISO noise is the digital equivalent of *grain* in digital cameras.

Step 3. Optical character recognition

In this step, the user identifies the historical period of the document and the OCR model based on which the document is recognized (*Figure 6*). An example of a recognized document is shown in *Figure 7*.

Step 4. Verification and editing of the recognized text

The text obtained in the previous step can be processed manually: this step integrates text editing modules such as virtual keyboards for different alphabets and spellchecks. An example is shown in *Figure 8*.

Step 5. Transliteration of the recognized and edited text

In this step, the transliteration tool from the old Romanian Cyrillic alphabet into the contemporary Latin alphabet is integrated. The transliteration is based on rules sets according to different epochs and alphabets (Romanian Cyrillic, Moldovan Cyrillic, transitional alphabets) and exception dictionaries. *Figure 9* provides an example, with recognized text on the left and its transliteration on the right.

Step 6. Verification and editing of the transliterated text

This step in which the transliterated text is verified is similar to step 4. Both the user and the spellchecker work with the same word dictionaries.

Step 7. Saving of the results

Together, all of the results, images, and texts that have been obtained from the input after steps 1 through 6 were completed, constitute a digitized document (see also *Figure 9*). This step also involves modules for downloading the preprocessed images, the recognized text, and the transliterated text and for uploading these texts and images to portal e-Moldova[13] (Caftanatov 2021).

To facilitate accessibility to the old documents we elaborated:

[13] https://emoldova.org/

- two OCR models trained on over 10,000 training examples and a dictionary of over 4,500 words for the 17th century;
- more than 10 OCR models with dictionaries containing more than 10,000 words for the 18th and 19th centuries;
- and an OCR model containing over 450,000 words for the 20th century.

The recognition of texts from the 18th and 19th centuries resulted in a WER (Word Error Rate) of 3-4.5%, and the WER for the 17th century was more than 6%.

4. Annotation and Conversion in Universal Dependencies

A qualitative corpus needs linguistic annotation. This section describes the work on morphosyntactic annotation of a corpus of old texts. Initially, the corpus building was started by Cenel Augusto Perez (Perez 2014) and called the Romanian Dependency Treebank (UAIC-RoDia-DepTb), or RoDia for short. In 2017, it obtained an International Standard Language Resource Number (ISLRN 156-635-615-024-0),[14] and its size was 16,000 sentences (or 320,000 tokens, including punctuation). Verification of this corpus's morphological and syntactic annotation was entirely manual. Since 2014, the corpus has been growing thanks to the ongoing work of Cătălina Mărănduc (Mărănduc 2022). In addition to being syntactically annotated, the corpus is presented in several formats: (1) the classic syntactic, containing more than 39,000 sentences, (2) the Universal Dependencies syntactic, and (3) a new syntactic-semantic format. To our knowledge, it is the largest syntactically annotated Romanian corpus where all of the verification was done only manually. Currently, the RoDia corpus contains annotated old texts and small volumes of non-standards texts, including regional and social media content, which enabled us to focus on the annotation of non-standard text types, such as oral regional fiction, social media communication, poetry, and old Romanian texts.

The creation of this corpus served several objectives:
(1) to attract attention to the linguistic diversity of the Romanian language;

[14] https://islrn.org/resources/156-635-615-024-0/

(2) to conduct a diachronic analysis of the Romanian language;
(3) to create a standard approach to annotating various types of Romanian texts;
(4) to build robust machine learning models for different types of annotations.

Old Romanian Cyrillic	**Modern Romanian**	**Morpho syntactic tag**	**Morphosyntactic tag description**
Їwсиф	Iosif	Npmsrn	Noun, personal, masculine, singular, direct case, non-articulated
фѳци	fugi	Vmis3s	Verb, main, indefinite, singular, 3rd person, subjunctive mode
кѳ	cu	Spsa	Preposition, accusative case
ЇС	Iisus	Npmsrn	Noun, personal, masculine, singular, direct case, non-articulated

Table 2: An example of text enriched with morphosyntactic tags and lemmas

4.1. Morphological annotation

The texts were automatically processed by the Ro-bin-hybrid Part-of-Speech (PoS) tagger (Simionescu 2011), using MULTEXT East project PoS tags (Erjavec 2012). This set of tags has been developed in unified mode for all European languages. Each tag is an abbreviation that starts with a upper-case letter, indicating the part of speech, and continues with a series of lower-case letters, which encode other morphological features dependent of the part of speech. For example, *Ncfpoy* deciphers as N=noun, c=common, f=feminine, p=plural, o=case (dative or genitive), and y=articulated.

FORM	**POSTAG**	**LEMMA**
anfivolie	Ncfsrn	amfivolie
anfivolii	Ncfprn	amfivolie
ăbăluiască	Vmsp3	ăbălui
palma	Ncfsry	palmă
sămînţa	Ncfsry	sămânță
partea	Ncfsry	parte

FORM	POSTAG	LEMMA
săvîrşi	Vmis3s	săvârși
săvîrşi	Vmn	săvârși

Table 3: A fragment of the lexicon for old Romanian texts, tagged by MULTEXT

Due to the rich morphology of the Romanian language, a set of as many as 614 morphosyntactic tags were developed for it during the MULTEXT project. This set was quite large and included a detailed description of the specifics of Romanian morphology. We simplified the tags, retaining 450 from the initial set and adding approximately 100 more tags to annotate specific elements of old Romanian. *Table 2* contains an example of text that has been enriched by the morphosyntactic tags and lemmas.

It is obvious that the accuracy of automated morphological tagging for this large set of tags is lower than it would be for a state-of-the-art equivalent used on the English language, for example. The best accuracy reported for the automated morphological tagging on modern Romanian texts and reduced tagset was 99% (Tufiș 2000); on old Romanian texts, it was considerably lower. To boost the accuracy for the non-standard texts, we enriched the dictionary for the Ro-bin-hybrid PoS-tagger with old Romanian words. *Table 3* contains several lexical units, each consisting of a word form, a word lemma, and a morphosyntactic description tag.

The first set of old texts was annotated using the tagger that had been trained on modern texts. Then, the texts were manually checked and corrected, and the tagger was trained on these corrected texts. Furthermore, a gold annotated corpus of non-standard texts was created through bootstrapping, whereby the processes of annotation, correction and re-training the tagger were repeated multiple times. Ultimately, the accuracy of PoS tagging for old texts reached 96% (Mărănduc 2022). It is extremely difficult to reach a higher level of accuracy, as each text that is subsequently added to the corpus differs considerably from those that have already been tagged, and the rate of words in that text that are unknown to the tagger is high, up to 12% of the new text. Each text that was added to the corpus was tagged in automate mode, and each tag was then checked and corrected manually. *Figure 10*

presents an excerpt of the tagger output: text enriched with morphosyntactic information coded in XML format.

4.2. Syntactic annotation

Initially, the corpus had been annotated in accordance with dependency parsing conventions that were developed at "Alexandru Ioan Cuza" University of Iași (UAIC), especially Functional Dependency Grammar (FDG), which features classical syntactic labels including numerous semantic sub-classifications of modifiers (Mărănduc 2017).

Dependency grammar is a class of modern grammatical theories that are based on the notion of dependency relation between linguistic units (that is, words) in a sentence. It means that words are connected to each other through directed links (Florea *et al.* 2014) forming a tree graph. The sentence predicate (verb) is considered the structural center of clause structure or a root of the graph. All other syntactic units (in this case, words) are connected to this verb or other words by means of links or dependencies forming branches and leaves of the graph. The main rule for this type of annotation is that every word except for the head (main verb) in a sentence has to be connected to only one other word, as a child is connected to a parent in a tree graph. The rules of connections vary depending on the corpus annotation guidelines. Usually, the type of connection is indicated by labels, such as "subject" or "modifier". The list of dependency types is limited and described in the annotation guidelines. For UAIC corpus annotation, 44 types of syntactic relations were defined; *Table 4* presents several examples of relation types.

Nr. crt.	**UAICtb syntactic relation**	English translation of relation	Example	English translation of example
1	a.adj.	adjectival attribute	fată *frumoasă*	*beautiful* girl
2	a.adv.	adverbial attribute	ziua *de ieri*	day *yesterday*
3	ap.	apposition	el, *autorul*	he, *the author*
4	a.pron.	attribute expressed by pronoun	cearta *cu el*	quarrel *with him*
5	a.subst.	attribute expressed by noun	cearta *cu tata*	quarrel *with Dad*
6	aux.	auxiliary verb	*am* terminat	*have* finished
7	a.vb.	verbal attribute	plăcerea *de a mânca*	the pleasure *of eating*
8	c.ag.	agent	bătut *de hoți*	beaten *by thieves*

Table 4: A sample of syntactic relation labels, with examples

The first texts of the corpus were annotated using a parser trained on the UAIC Romanian Dependency Treebank (UAICRoDepTb), and the automate annotation had to be verified and corrected manually because its accuracy on the non-standard old texts was quite poor. Then, MaltParser[15] was trained on the verified texts, and new portions of the corpus were annotated using its various models. Every text, annotated by a parser was checked and corrected manually in order to create a gold standard corpus for further parser training.

Figure 11 presents RoDia in XML format, where the XML tag attributes *head* and *deprel* define the syntactic relation. The *head* value indicates the current word head; the *deprel* value specifies the label of the relation between the words. Each word is indexed within the sentence and has a unique identifier ("id") that is used to indicate the head. For example, the word with id=4 has head=1, which means that this word is connected to the word with id=1. The attribute *deprel*=c.c.l. means that, in relation to the head, this word has the role of *complement circumstanțial de loc* (adverbial phrase of place). The attribute *form* contains the initial form of the word as it appears in the text, and the attribute *postag* indicates the word's morphosyntactic description. In XML markup, it is possible to observe relations between words, but verifying and correcting the annotation in this format is quite difficult. In order to facilitate the work of linguists, we have developed a special graphical annotation tool called TreeAnnotator (Moruz 2008).

This graphical annotator has many advantages. For one, its visual design is user-friendly: the sentence is presented as a tree, where the head is located higher than its descendants, and where each of the other words in the sentence is presented as a branch stemming from the head and is positioned on the same vertical plane as the corresponding word listed in the version of the sentence that is presented along the bottom of the image. Another advantage is that the tree is completely editable. Any connection can be removed, added, redirected or renamed. All this can be done with the mouse, or through the use of a combination of keys, which increases the speed of work, as the mouse can be used to select a word or a relationship, while the left hand

[15] http://www.maltparser.org/

is used to push the keys. However, the labels are introduced manually, which increases the risk of typing errors. Still, the greatest drawback of this Graphical User Interface (GUI) is the fact that its software is no longer supported by its developper. The code can no longer be modified, which means that it can only be used for the old format of the UAIC treebank.

Figure 12 presents an example of a syntactic tree that has been visualized graphically using TreeAnnotator. The sentence is listed below the tree, and morphosyntactic tags (which are also editable) are presented below the words of the sentence. The menu at the top contains buttons with the editing options.

Currently, the corpus contains a small amount of contemporary texts, including social media content. A large part of the corpus consists of old texts from the 16th–19th centuries. The initial intention had been to annotate a larger number of language styles, in rather equal proportion; once we added the Alba Iulia New Testament (1645) to the corpus, however, the ecclesiastical style became the most represented. This is quite normal for that period, when written texts were predominantly found in churches. We recognize that the church style in our corpus is overrepresented to the detriment of other styles though it was not our intention. Subsequently, several other historical works were integrated into the corpus. These include a book containing moral philosophical sentences and narrative examples, *Flower of Gifts*, 1592 (*see* Gheţie and Mareş 1996); a book containing laws established by Prince Caragea in 1818; Neculce's *Chronicles of the Land of Moldavia* (1745); and Dosoftei's *Psalms*. Currently, the texts that have been annotated using the UAIC convention represent 34,794 sentences (714,377 tokens) and reflect an average of 20.53 words per sentence.

4.3. Universal Dependency Annotation

In order to widen the visibility of the corpus, its annotation has been transformed into the international Universal Dependencies project convention. Universal Dependencies[16] is a framework for the consistent annotation of grammar (parts of speech, morphological features, and syntactic dependencies) across different human

[16] https://universaldependencies.org/

languages. UD is an open community effort with over 500 contributors producing over 200 treebanks in over 100 languages (Marneffe *et al.* 2021).

The principles of the UAIC treebank annotation, which were established in 2007 and initially intended for didactic use, are quite different from those of the UD. UD annotation convention highlights meaningful words (such as nouns, verbs, adjectives) and subordinates functional words (such as conjunctions, prepositions, articles) to them. In the initially created UAIC corpus, the types of relations are highlighted, and the functional words are considered connecting nodes, thus they co-relate the head and the subordinated meaningful word in the tree. The main difference between these two systems of annotation lies in the annotation of functional words, such as copulative verbs: they are considered heads in the UAIC treebank, whereas they are subordinated in UD. Another difference is in the number and types of dependency labels. The UAIC treebank annotation uses the same tag to mark a relation expressed by a word as it does to mark a relation expressed by a subordinate clause. By contrast, UD uses a range of different tags, such as *ccomp*, *csubj* or *advcl* (where the initial *c* or the final *cl* means *clause*).

Figure 13 compares UAIC and UD connection principles. In the UAIC example, the complement "în ceriu" (to heaven) is connected to "sui" (ascends) as a *complement circumstanțial de loc* (adverbial phrase of place), where the preposition "în" is the head of "ceriu". In the UD example, the type of the complement is not specified, and the preposition "în" is subordinated to the more important word "ceriu."

UAIC and UD also differ in terms of the number and naming of syntactic relation types. UAIC identifies 14 kinds of adverbial modifiers, thus offers rich semantic information; however, UD proposes 65 syntactic relation types. *Table 5* presents several examples of how UD and UAIC syntactic relation types correspond to one another. As shown, there are some cases where a given UAIC relation type has an exact correspondence with a UD relation type, but there are also cases where a UAIC relation type corresponds to two or more UD relation types.

UD syntactic relation types	**UAIC syntactic relation types**	**Explanations**
amod, det, nummod	a.adj.	adjectival attribute

advmod	a.adv.	adverbial attribute
nsubj, csubj, nsubj:pass, csubj:pass	sbj.	subject
nmod:pmod, ccomp:pmod	c.prep	prepositional complement

Table 5: correspondences between UD and UAIC tags for syntactic relation types

As a result of these discrepancies, serious problems arose during the process of transformation of the syntactic annotation of the existing UAIC corpus into UD convention/ protocol of annotation. After detailed analysis, we created a rule-based transformation tool, called TREEOPS, to convert UAIC annotation into UD annotation (Colhon *et al.* 2017). The tool works in the XML format, and the rules follow an if-then (condition-action) logic. There are three types of rules:

- rules that have only one condition;
- rules that have two or more conditions (and actions); and
- rules that are used for structural transformations (such as reversing the direction of the relation between the noun and the prepositional case marker).

There is also a list of super-rules that determine the order in which the rules described above should be applied. There are more than 150 rules for the transformation into the UD convention. For example, for the transformation of an adverbial in the UD convention, there are three rules for each type of adverbials, as identified and explained below. Belonging to the second type of rules, these rules have two conditions: the condition of a dependency relation and that of a morphological analysis.

1. //word[@deprel='c.c.l.' and starts-with(@postag, 'R')]/@deprel => changeAttrValue('advmod')

- **If the word** has a dependency relation label c.c.l. and its morphological tag starts with R (which means it is an adverb), then the dependency relation label should be changed to advmod.

2. //word[@deprel='c.c.l.' and starts-with(@postag, 'V')]/@deprel => changeAttrValue('advcl')

- **If the word** has a dependency relation label *c.c.l.* and its morphological tag starts with *V* (which means it is verb), then the dependency relation label should be changed to *advcl*.

3. //word[@deprel='c.c.l.' and (starts-with(@postag, 'N') or starts-with(@postag, 'P') or starts-with(@postag, 'M'))]/@deprel => changeAttr Value('obl').

- **If the word** has a dependency relation label *c.c.l.* and its morphological tag starts with *N* (which means it is a noun), then the dependency relation label should be changed to *obl*.

The transformation errors are often caused by unexpected syntactic structures commonly encountered in a natural language, especially in a nonstandard one. Monitoring the output of TREEOPS is therefore absolutely necessary, although the task is not as time-consuming as that of correcting the morphological and syntactic output of the parser and PoS-tagger.

The final volume of the Romanian Nonstandard UD corpus is smaller than the initial RoDia corpus; all of the initial UAIC annotations that were converted into UD need to be verified manually, which is a slow process. A number of the transformed texts still needs to be verified and corrected.

4.4. Semantic Annotation

While UD version of the corpus improves the visibility of our corpus within the community, the UAIC annotation model is better tailored to the syntactic particularities of the Romanian language, and we still annotate in this format to preserve all the information that has been automatically annotated and carefully monitored in the initial corpus. However, syntactic annotation, even when it is detailed, does not code the semantic content of the sentences, and it therefore proved necessary to add a semantic layer to the classical treebank.

Used to highlight the semantic information present in syntactic and morphological UAIC-RoDepTb annotations (Mărănduc *et al.* 2017), the logical-semantic format has been inspired by the Tectogrammatic layer of the Prague Dependency Treebank (PDT) (Bohmová *et al.* 2003) and Abstract Meaning

Representation (AMR) (Bănărescu *et al.* 2013). AMR semantic annotation is not a dependency tree, and rather than referring to words, its nodes consist of concepts; AMR therefore could not be adopted as it is for our semantic annotation. The principles underlying the Tectogrammatic layer of the PDT correspond much more closely to our own. Furthermore, it includes categories for annotating semantic information found in exclamatory and interrogative sentence forms, and in verb moods and tenses, by considering the punctuation or morphological annotation as containing semantic information.

Enriching the syntactic annotation, we identified 96 semantic labels for semantic dependency relations, as summarized below (for the full list of semantic labels, *see* Mărănduc 2022).

- Adverbial relation labels indicate:
    - semantic labels for the relations of temporality and spatiality (tagged as TEMP, PAST, FTR, LOC);
    - logical relationships: cause (CAUS), purpose (PURP), consequence (CSQ), opposition (OPPOS), concession (CNCS), condition (COND), exception (EXCP), cumulation (CUMUL), association (ASSOC), reference (REFR), restriction (RESTR), result (RSLT).
- Information about the names of objects or actors is derived from the classification of the pronouns: appurtenance (APP), deictic (DX), emphatic (IDENT), undefined (UNDEF).
- The classification of functional words (which are considered to be connectors) provides the classification of relations.
    - Four of them are coordinating connectors: inclusion (CNADVS), reunion (CNCNCL), disjunction (CNDISJ), and intersection (CNCONJ).
    - A fifth is the subordinating connector (CNSBRD).
    - A sixth is the copulative verb “to be” (CNCOP) and serves to connect the subject and its description, as rendered by the predicate noun.

- Another set of labels conveys information regarding the reality of the action: optative (OPTV), uncertain (UNCTN), potentiality (POTN), generic (GNR), dynamic (DYN).
- Five quantifiers are considered: necessity (QNECES), possibility (QPOSSIB), existence (QEXIST), universality (QUNIV), and negation (QNEG).
- Another set of labels provides information about the communication in text: addressee (ADDR), blame (BLAM), greeting (GREET), politeness (POLIT), interrogative (INTROG), exclamatory (EXCL), incidence (INCID).

TREEOPS—the tool described in the previous section and used to convert UAIC annotation into UD—was also used to convert the initial UAIC format into the semantic one. In this transformation, the dependency connections remained intact, but each of the 44 syntactic labels had to be replaced with one of the 96 semantic labels. Twenty-one syntactic relations were semantically monovalent; each of them has exactly one corresponding semantic label. The others are semantically ambiguous, and specific rules of transformation have been created for them. For example, the syntactic tag *auxiliary* (AUX) can be transposed into OPTV or PAST, if the form of the auxiliary (and of the conjugate verb) indicates the conditional mode or the past tense, or if the auxiliary indicates the future tense (FTR) or passive (PASS).

Format	Number of sentences	Number of tokens
UAIC syntactic (RoDia)	32,753	671,235
UD syntactic (UD Non-Standard)	16,936	348,562
UAIC semantic	5,566	99,341

Table 6: Size of three treebank formats

Syntactic relations that are labeled according to morphological criteria—such as *a.subst.* (noun attribute), *a.vb.* (verbal attribute), and *c.prep.* (prepositional complement)—have high semantic ambiguity; they can have almost any semantic value. The automatic tool leaves such labels untouched, and they are edited manually. As a result, the part of the corpus that is semantically labeled remains relatively small, and its expansion requires manual work by linguists. When a large enough portion of the corpus has undergone this type of annotation, machine learning parsers trained on it are able to produce semantically labelled text with high accuracy and manual work will be reduced considerably (Colhon *et al.* 2017).

4.5. Automate Parsing Evaluation

The RoDia Dependency TreeBank and Nonstandard Universal Dependencies corpus were created using the bootstrapping approach, meaning that, after undergoing automated annotation, which involved various applications, the treebank and corpus were manually verified and corrected. The sizes of each of the verified and corrected corpora are presented in *Table 6*. The corpora annotated syntactically are large enough to be used for machine learning systems training. In this section, we report on results from the parsing experiments we conducted using MaltParser, one of the tools used in the UD project for automated dependency parsing.

Book Entry	**Sub-corpus**	**# of sentences**	**Words per sentence**	**UAIC UAS**	**UAIC LAS**	**UD UAS**	**UD LAS**
1	apostles	5,894	20.86	83.76	76.29	87.42	82.99
2	Caragea.Law	1,188	23.95	73.34	65.31	83.62	79.09
3	Flower.gifts	1,088	19.46	83.24	74.72	87.56	83.69
4	Folk.Mol	1,806	18.69	80.04	72.37	86.75	82.22
5	Folk.Rom	695	23.61	71.85	64.5	86.08	78.29
6	gospel	5,172	17.86	85.11	78.66	84.06	76.47
Average			20.74	**79.56**	**71.98**	**85.92**	**80.46**

Table 7: The accuracy of automated parsing on different parts of the UAIC RoDia corpus and for the UD format of the same parts of the corpus

MaltParser implements a transition-based approach to dependency parsing, which means that it consists of a transition system for deriving dependency trees and

a classifier for deterministically predicting the next transition based on a feature representation of the current parser configuration (Nivre *et al.* 2007). MaltParser is a complex system with many parameters that need to be optimized and adapted to the task at hand. *Table 7* presents the obtained accuracy for several parts of the UAIC syntactic corpus when various MaltParser algorithms are used. Parsing accuracy is measured with two classical metrics. On the one hand, the Labeled Attachment Score (LAS) reflects the ratio between the number of correct attachments with correct labels and the total number of attachments. On the other hand, the Unlabeled Attachment Score (UAS) indicates the ratio between the number of correct attachments and the total number of attachments, where labels are not taken into consideration. The last two columns of *Table 7* present the same metrics for the UD format of the same corpus parts.

As the table shows, automated parsing for the UD format is better than for its UAIC counterpart. This is probably because MaltParser was created for the UD project and is better able to detect UD principles than to recognize those of the UAIC mode of dependency attachments.

5. Conclusion

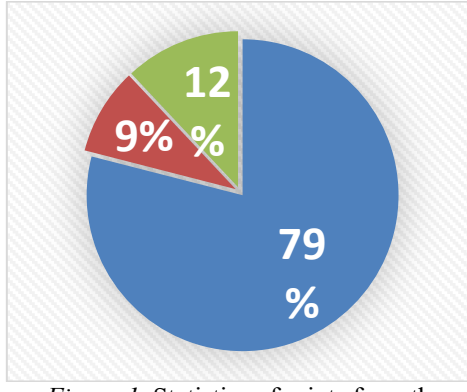
Our paper introduces a digitization platform[17] that offers access to a comprehensive set of tools and informational resources, specifically designed for digitizing Romanian documents printed in the Cyrillic script spanning the 17th to 20th centuries. We also take pride in our significant contribution to the development and enrichment of an annotated Romanian corpus, encompassing a diverse range of old texts, regional folklore, and nonstandard texts from Moldova and Romania.

The digitization of these old documents presents captivating scientific challenges, and since Natural Language Processing (NLP) for such texts is typically the purview of isolated and scattered research groups, the resulting products often lack uniformity in terms of formats and standards. Through our ongoing efforts, we are dedicated to digitizating old Romanian texts, facilitating their transliteration into modern Romanian (Latin) alphabet, enhancing optical character recognition, adding

[17] https://digitizare.math.md.

annotations, and ultimately incorporating these old texts into the Universal Dependency (UD) repository, adhering to a common standard (Bobicev *et al.* 2019). This initiative not only preserves a valuable piece of cultural heritage but also fosters collaboration and cohesion in the field of NLP, ensuring that a wider community of scholars and enthusiasts have access to these resources.

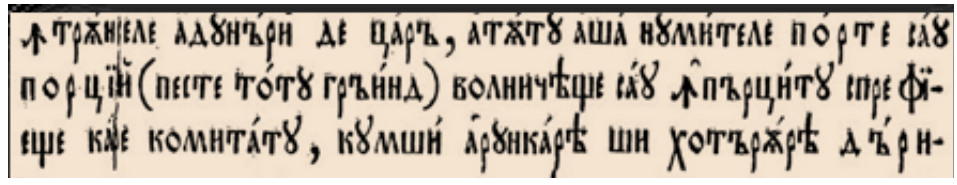
Annex 1: Figures



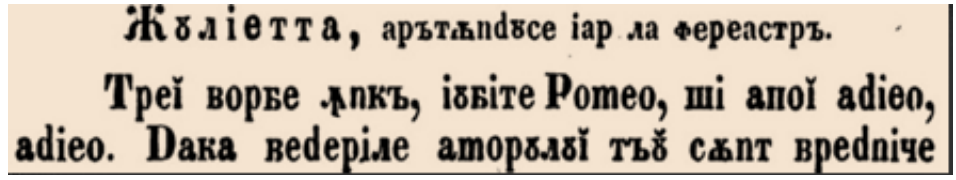
79% Romanian
9% Greek
12% Other (Bulgarian, English, French, German, etc.)

Figure 1: Statistics of prints from the years 1508-1830 at the Library of the Romanian Academy.

Sample of 18th century typographic script



Sample of 19th century typographic script (mixed/ transitional alphabet)



Sample of 20th century typographic script

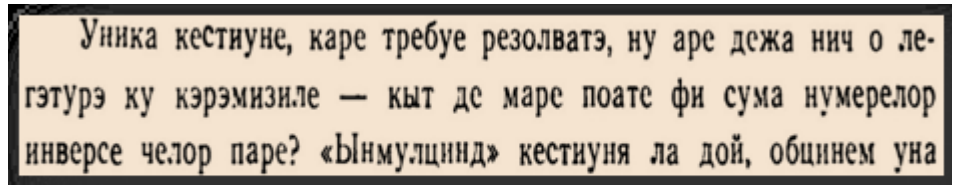


Figure 2: Samples of texts from different centuries.

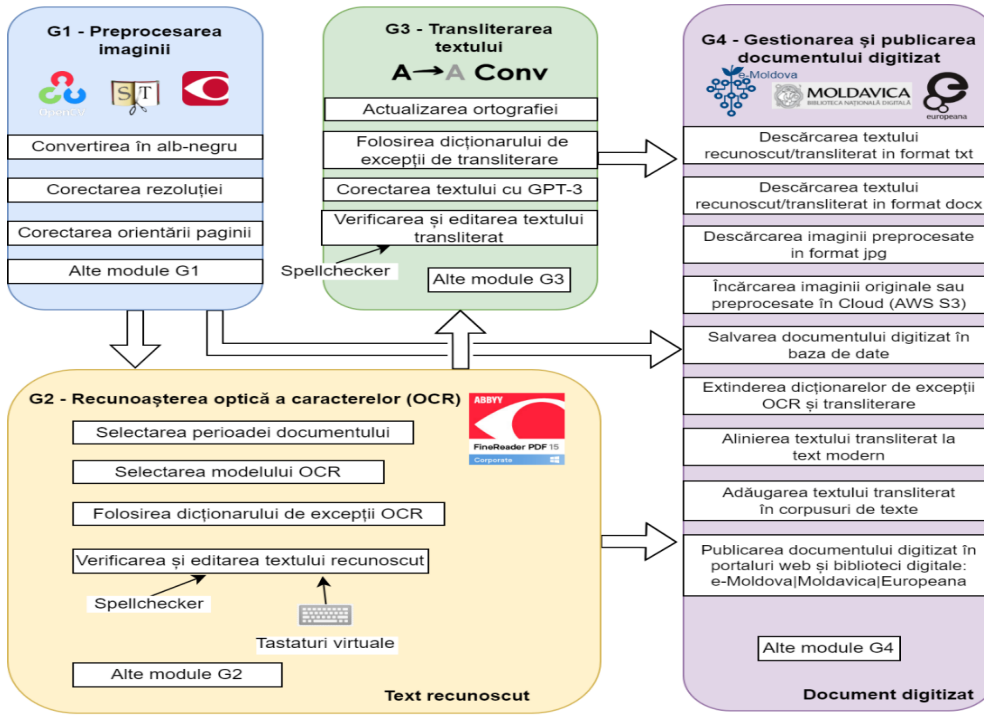
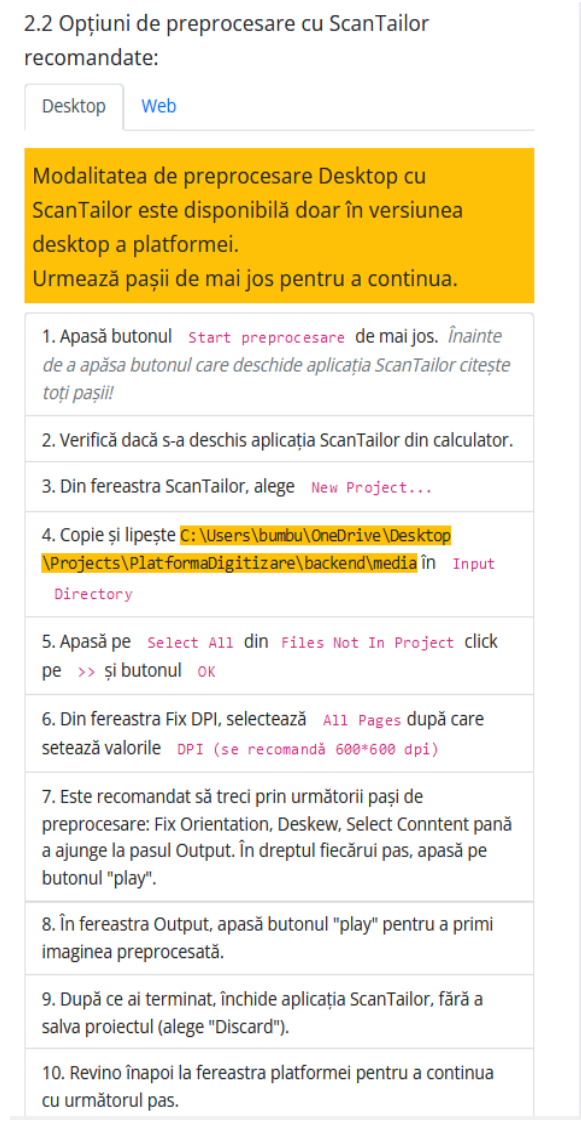


Figure 3: The architecture of Digitization Platform (DP)



2.2 Opțiuni de preprocesare cu ScanTailor recomandate:

Desktop | Web

Modalitatea de preprocesare Desktop cu ScanTailor este disponibilă doar în versiunea desktop a platformei.
Urmează pașii de mai jos pentru a continua.

1. Apasă butonul `Start preprocesare` de mai jos. *Înainte de a apăsa butonul care deschide aplicația ScanTailor citește toți pașii!*
2. Verifică dacă s-a deschis aplicația ScanTailor din calculator.
3. Din fereastra ScanTailor, alege `New Project...`
4. Copie și lipește `C:\Users\bumbu\OneDrive\Desktop\Projects\PlatformaDigitizare\backend\media` în `Input Directory`
5. Apasă pe `Select All` din `Files Not In Project` click pe `>>` și butonul `OK`
6. Din fereastra Fix DPI, selectează `All Pages` după care setează valorile `DPI (se recomandă 600*600 dpi)`
7. Este recomandat să treci prin următorii pași de preprocesare: Fix Orientation, Deskew, Select Conntent pană a ajunge la pasul Output. În dreptul fiecărui pas, apasă pe butonul "play".
8. În fereastra Output, apasă butonul "play" pentru a primi imaginea preprocesată.
9. După ce ai terminat, închide aplicația ScanTailor, fără a salva proiectul (alege "Discard").
10. Revino înapoi la fereastra platformei pentru a continua cu următorul pas.

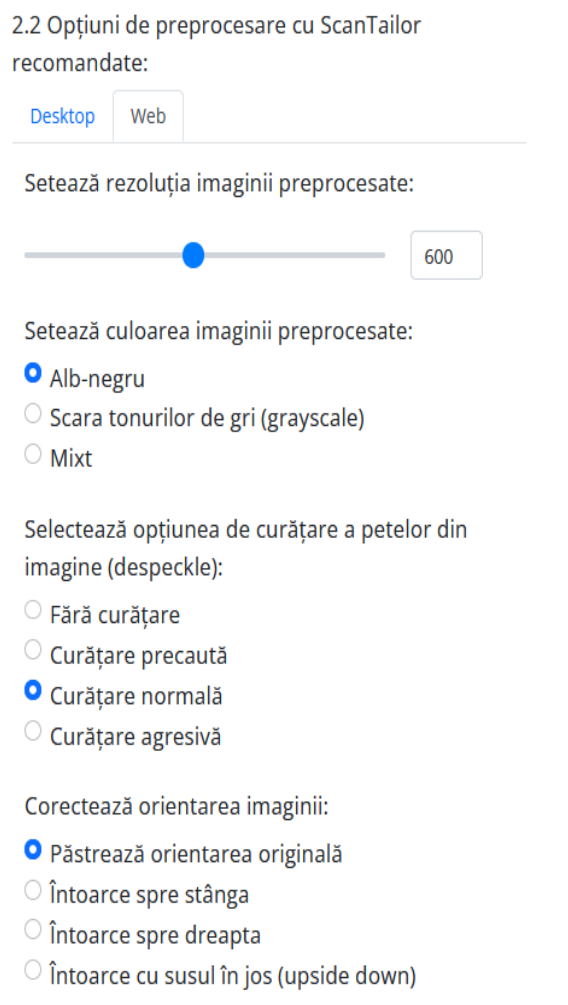
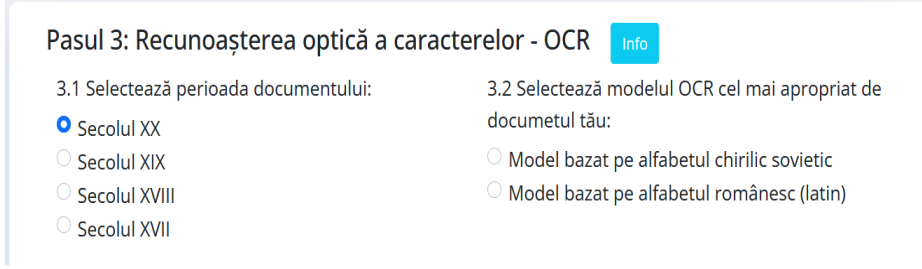


Figure 4: Two ways of preprocessing with Scan Tailor: desktop preprocessing (left), web preprocessing options (right).



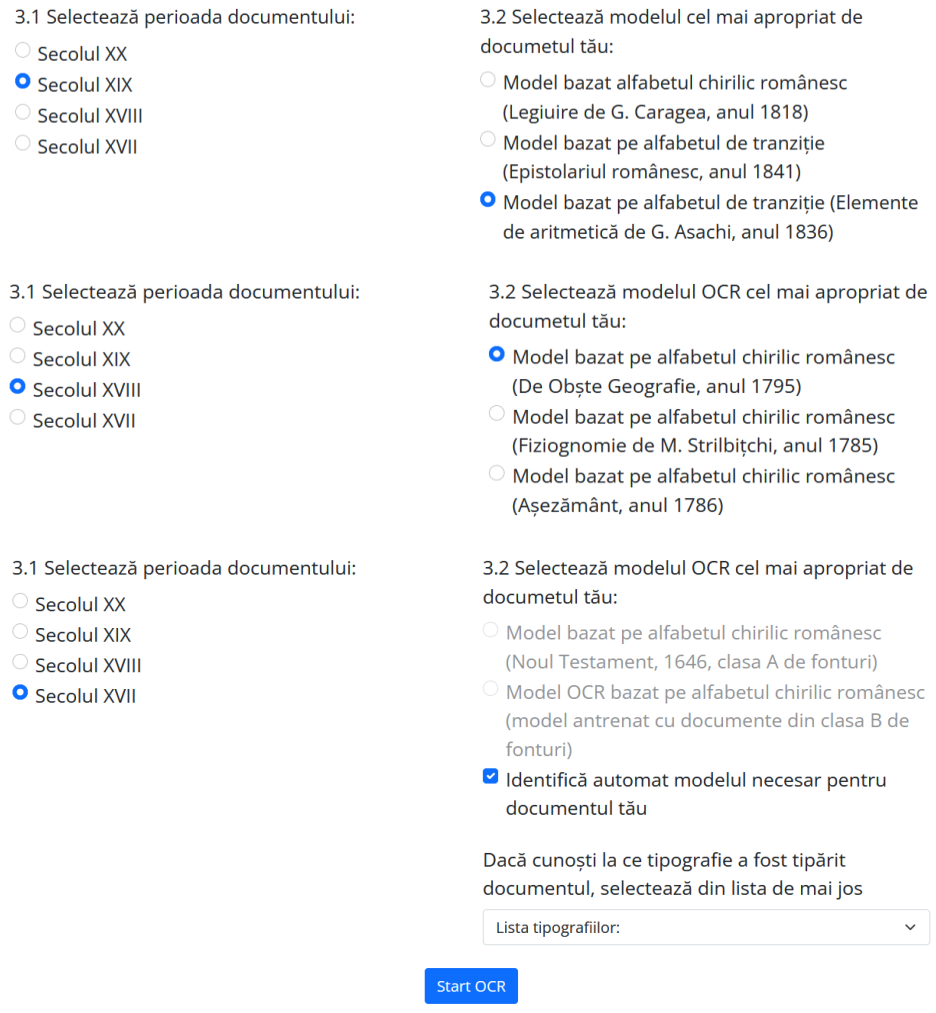


Figure 5: Options available in step 3.

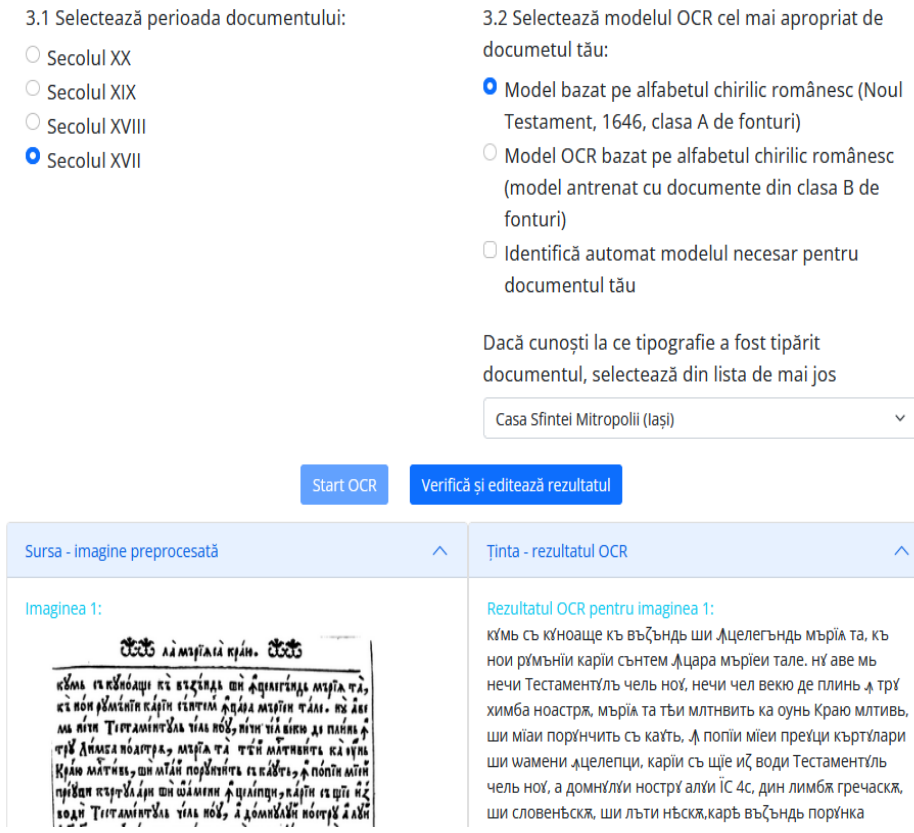


Figure 6: Running step 3 in the digitizing application.



Figure 7: Verification and editing of the recognized text: recognized text (left), processed image (right), virtual keyboard (bottom).

кѹмь съ кѹноаще къ възъндь ши ѧцелегъндь мъріѧ та, къ нои рѹмънїи карїи сънтем ѧцара мъріеи тале. нѵ аве мь нечи Тестаментѹлъ чель ноѵ, нечи чел векю де плинь ѧ трѹ химба ноастрѫ, мъріѧ та тѣи млтнвить ка оунь Краю млтивь, ши мїаи	cum să cunoaște că văzând și înțelegând măria ta, că noi rumânii carii sântem în țara măriei tale. nu ave m neci Testamentul cel nou, neci cel vechiu de plin întru limba noastră, măria ta te-ai milostivit ca un Craiu milostiv, și mi-ai poruncit să caut, în popii miei preuți cărtulari și

Figure 8: Transliteration of a recognized text (Cyrillic on the left side and Latin on the right side).

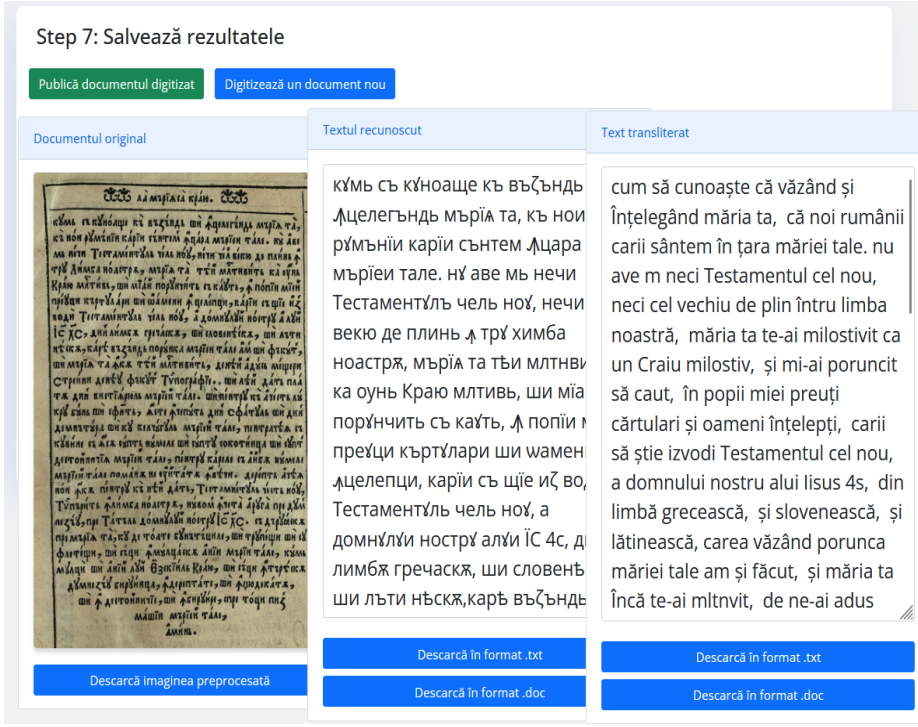


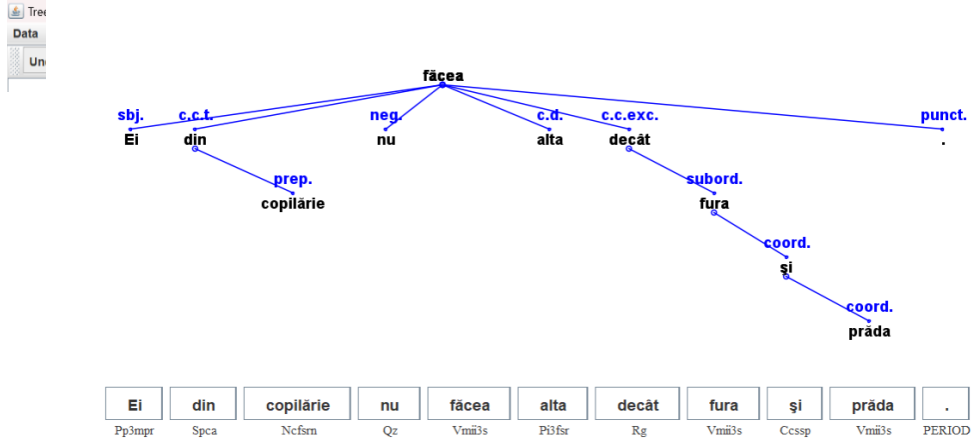
Figure 9: Showing the original document (left), the recognized text (middle), and the transliterated text (right).

```
<POS_Output>
 <S id="1" offset="0">
  <W Case="direct" Definiteness="no" Gender="masculine"
LEMMA="bun" MSD="Afpmsrn" Number="singular"
POS="ADJECTIVE" id="1.1" offset="0">Bun</W>
  <W Case="direct" Definiteness="no" Gender="masculine"
LEMMA="cal" MSD="Ncmsrn" Number="singular" POS="NOUN"
Type="common" id="1.2" offset="4">cal</W>
  <W EXTRA="tranzitiv" LEMMA="hrăni" MSD="Vmip3s"
Mood="indicative" Number="singular" POS="VERB" Person="third"
```

Figure 10: XML format of the tagger (W=Word, MSD=MorphoSyntactic Description). The initial word form, the lemma, all morphological features of the word are coded as separate attributes of the tag W.

```
<sentence id="20" parser="" user="ugla" date="2014-14-15">
  <word id="1" form="Era" lemma="fi" postag="Vmii3s" head="0" />
  <word id="2" form="una" lemma="unul" postag="Pi3fsr" head="1"
deprel="sbj."/>
  <word id="3" form="şi" lemma="şi" postag="Rg" head="4" deprel="a.adv."/>
  <word id="4" form="pe" lemma="pe" postag="Spsa" head="1"
deprel="c.c.l."/>
  <word id="5" form="faţada" lemma="faţadă" postag="Ncfsry"
```

Figure 11: An example of the XML format of RoDia.
Each word is presented in a tag "word" that contains the word and all its annotated attributes.



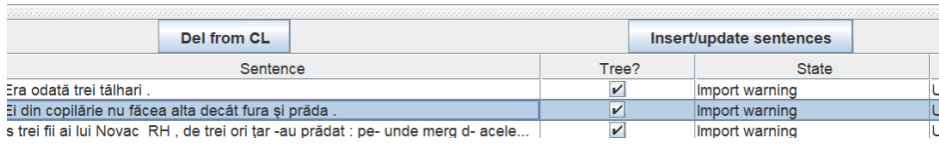


Figure 12: TreeAnnotator interface.

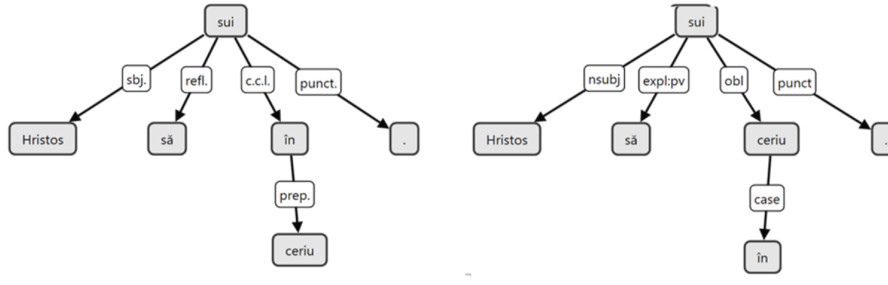


Figure 13: Comparison of UAIC (left) and UD (right) connection principles.
The sentence is: *Hristos să sui in ceriu* [Hristos ascends to heaven]

References

Bănărescu, Laura, Claire Bonial, Shu Cai, Mădălina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philip Koehn, Martha Palmer, and Nathan Schneider. 2013. "Abstract Meaning Representation for Sembanking." In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 178–186.

Bobicev, Victoria, Maranduc Cătălina, Tudor Bumbu, Ludmila Malahov, Alexandru Colesnicov, and Svetlana Cojocaru. 2019. "Contribution to the Universal Dependencies Treebank of Non-Standard Romanian Texts." In *Proceedings of the Language Technologies for All (LT4All)*, Paris, UNESCO Headquarters, France, 5–6 December, 300–303.

Bohmová, Alena, Jan Hajič, Eva Hajičova, and Barbora Hladka. 2003. "The Prague Dependency Treebank: A Three-Level Annotation Scenario. Text, Speech and Language Technology." Berlin and New York: Springer.

Boian, Elena, Constantin Ciubotaru, Svetlana Cojocaru, Alexandru Colesnicov, and Ludmila Malahov. 2014. "Digitizarea, Recunoaşterea și Conservarea Patrimoniului Cultural-Istoric." *Akademos* 1(32): 61–68.

Bumbu, Tudor. 2021. "Towards a Font Classification Model for Romanian Cyrillic Documents." *Computer Science Journal of Moldova* 3(87): 291–298.

Bumbu, Tudor. 2023. "Tehnologii și Resurse Informaționale pentru Digitizarea și Procesarea Textelor din Patrimoniul Historico-Cultural." PhD diss., State University of Moldova.

Caftanatov, Olesea, Tudor Bumbu, Lucia Erhan, Iulian Cernei, Veronica Iamandi, Vasile Lupan, Daniela Caganovschi, and Mihail Curmei. 2021. "Discover the Moldovan Cultural Heritage through e-Moldova Portal by Using Crowdsourcing Concept." In *Proceedings of the Workshop on Intelligent Information Systems (WIIS2021)*, Chișinau, Republic of Moldova, October 14–15, 65–75.

Cazimir, Ștefan. 2006. *Alfabetul de Tranziție*. Bucharest: Humanitas.

Ciubotaru, Constantin, Svetlana Cojocaru, Alexandru Colesnicov, Valentina Demidov, and Ludmila Malahov. 2015. "Regeneration of Cultural Heritage: Problems Related to Moldavian Cyrillic Alphabet." In *Proceedings of International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, 177–184.

Cojocaru, Svetlana, Alexandru Colesnicov, and Ludmila Malahov. 2017. "Digitization of Old Romanian Texts Printed in the Cyrillic Script." In *ACM International Conference Proceeding Series*, 143–148.

Colesnicov, Alexandru, Ludmila Malahov, Svetlana Cojocaru, and Ludmila Burtseva. 2021. "Digitization Technology of Old Romanian Documents Printed in the Cyrillic Script." In *Horizons in Computer Science Research*. Volume 21, edited by Thomas S. Clary, 185–217. New York: Nova Science Publishers.

Colhon, Mihaela, Cătălina Mărănduc, and Cătălin Mititelu. 2017. "A Multiform Balanced Dependency Treebank for Romanian." In *Proceedings of Knowledge Resources for the Socio-Economic Sciences and Humanities*, Varna, Bulgaria, 9–18.

Cristea, Dan, Daniela Gîfu, Svetlana Cojocaru, Alexandru Colesnicov, Ludmila Malahov, Mihai Popescu, Mihaela Onofrei, and Cecilia Bolea. 2018. "How to Find Out Whether the Romanian Language Was Influenced by the Two Historical Unions?" In *Proceedings of 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language*, Iași, November 22–23, 77–87.

Erjavec, Tomaž. 2012. "MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages." *Language Resources and Evaluation* 46 (1): 131–142. Berlin and New York: Springer.

Florea, Iulia Maria, Traian Rebedea, and Costin-Gabriel Chiru. 2014. "Parser de Dependenţe pentru Limba Română Realizat pe Baza Parserelor pentru Alte Limbi Romanice." *Revista Română de Interacţiune Om-Calculator* 7 (1): 1–20.

Gheţie, Ion, and Alexandru Mareş (eds). 1996. *Cele mai Vechi Cărţi Populare în Literatura Română. Vol. 1: Floarea Darurilor. Sindipa.* Bucharest: Minerva Publishing House.

Mărănduc, Cătălina. 2022. "Instruments for Processing the Old Romanian Language (Training Corpora, Models, Statistics)." PhD diss., "Alexandru Ioan Cuza" University of Iași, Romania.

Mărănduc, Cătălina, F. Hociung, and V. Bobicev. 2017. "Treebank Annotator for Multiple Formats and Conventions." In *The 4th Conference of Mathematical Society of the Republic of Moldova (CMSM4)*.

Mărănduc, Cătălina, Monica Mihaela Rizea, and Dan Cristea. 2017. "Mapping Dependency Relations onto Semantic Categories." In *Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*.

Marneffe, Marie-Catherine, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. "Universal Dependencies." *Computational Linguistics* 47 (2): 255–308.

Moruz, Alexandru. 2008. "Dezvoltarea unui Adnotator FDG (Functional Dependency Grammar) pentru Limba Română." Master's diss., Facultatea de Informatică, Universitatea „Alexandru Ioan Cuza" Iaşi.

Nivre, Joakim, Johan Hall, Jens Nilsson, and Atanas Chanev. 2007. "MaltParser: A Language-Independent System for Data-Driven Dependency Parsing." *Natural Language Engineering* 13: 95–135. Cambridge: Cambridge University Press.

Perez, Cenel-Augusto. 2014. "Linguistic Resources for Natural Language Processing." PhD diss., Al. I. Cuza University, Iași.

Simionescu, Radu. 2011. "Hybrid POS Tagger." In *Workshop on Language Resources and Tools in Industrial Applications, Eurolan 2011 Summer School*.

Tufiș, Dan. 2000. "Using a Large Set of EAGLES-Compliant Morpho-Syntactic Descriptors as a Tagset for Probabilistic Tagging." *LREC*.

Digitalization of Romanian Lexicography

Elena Isabelle TAMBA

Romanian Language and the Digitalization Process

The digital age has brought significant transformations to Romanian linguistic and lexicographic research, providing opportunities for greater accessibility, collaboration and innovation in meeting the needs of language users. Philological research in Romania is characterized by an extensive digitalization process, which allows the Romanian language and, implicitly, Romanian culture to be promoted and preserved. This text is an overview of the state of digitization of lexicographic research in Romania. However, Romania is still in need of a coherent national plan to develop medium- and long-term strategies for digitizing the Romanian language.

Since language is both an expression of culture and an essential element of identity (whether individual or collective), as well as a means of communication, there is a constant need to adapt it to changes in human society. Thus, since the European Union is a multilingual society, its language policy has been defined based on respect for linguistic diversity and the creation of intercultural dialogue throughout its territory and through the educational component, which is a fundamental European value. The European Parliament resolution of September 11, 2018 on the state of languages in the digital age urges Member States to propose effective ways of supporting and preserving their own language(s) in the digital age, such as by developing databases and translation technologies in all EU languages, including those that are less widely spoken. In keeping with European language policy and strategies, Romania has also been engaged in recent years in exploring an extensive and comprehensive digitization process, including the creation of digital resources and tools in the field of philology, to facilitate access to resources geared towards the Romanian language.

Digitalization processes involve converting existing print [or: paper] resources to an electronic format (retro-digitization) and developing lexicographic or linguistic resources in an electronic format, using tools such as writing programs and sample extraction programs (to cite two examples from lexicography). The digital age brings a number of challenges as well as advantages for the field of philology, since digitization enables a change in the way resources and tools are created, while at the same time having implications for the way it is received by the public.

Romanian Digital Lexicography: a bird eye's view

In addition to affecting philological research in Romania, transformations in the digital age also have an impact on the country's lexicographic field. In this paper, we will present an overview of the main digital lexicographic projects carried out in Romania over the last few years.[1] These projects can be organized into two groups:

a. Academic projects: most efforts in the digitalization of lexicographic research have been made under the auspices of the Romanian Language institutes and the IT institutes of the Romanian Academy; a few projects are being developed in certain libraries and in research centers at certain universities in Romania.

b. Private projects: examples include dexonline (https://dexonline.ro/)—a lexicographic platform initiated by volunteers—and projects organized by certain publishing houses.

Digital Lexicography in the Romanian Academy: Lexicography within the Romanian academy has shifted significantly in the digital age in response to related changes in the way linguistic and lexicographic resources are developed, accessed and used. In particular, the new digital era has brought about innovative approaches to linguistic research in Romanian and the creation of Romanian-language resources, effectively transforming lexicographic methodologies. For instance, whereas dictionaries had previously been created based on examples that were identified

[1] For a larger presentation of the evolution of the digitalization of the language research in Romania, *see* Haja, Tamba, Teodorescu 2023; Tamba 2023; Tamba 2024.

manually in various books, we now use tools such as dictionary editing programs and engage in tasks such as contextual corpus searching using special programs.

Lexicographic research is primarily conducted by specialists, although non-specialists or volunteers can and do participate in certain aspects of this work. For example, dictionaries and lexicographic corpora are typically created by lexicographers, but members of the general public can also contribute to the development of certain lexicographic resources.[2] Researchers now have access to various digital tools and methodologies, including documentation from textual digital databases that contain data such as written texts, oral information transcribed into written form, and audio information. They use example extraction programs; dictionary writing systems; and other relevant technologies (cf. Haja *et al.* 2020). Romanian lexicographers are engaged in the creation various types of resources, such as different kinds of dictionaries, dictionary networks, and tools and resources for lexicography and linguistics.

The history of digitization projects in Romanian lexicography has been and is sometimes still complicated: whereas some projects are very active, others were abandoned for a while and later resumed, while others still have been progressing steadily but at a relatively slow pace. Nonetheless, the digitalization of research represents a special stage in the evolution of Romanian lexicography. On the one hand, digitalization involves obtaining the digital format of existing resources; on the other hand, it involves the creation of dictionaries, new resources, and instruments in an electronic format. At the same time, linguistic and lexicographic resources continue to be created in classical, printed format (Tamba 2023, 91).

Thesaurus Dictionary of the Romanian Language in the Digital Age: The main Romanian lexicographic project edited by the Romanian Academy is *Dicționarul limbii române* (DA/DLR) [Thesaurus Dictionary of the Romanian Language], which was started 115 years ago. Creating an electronic version of the dictionary that is

[2] Citizen scholars may engage directly in collaborative activities (situation common for many languages, if we are thinking about the Wiki projects), such as documentation, creation, development and testing of various resources. For a novel perspective on the involvement of laypersons in the documentation process for creating the first edition of the *Oxford English Dictionary* in the nineteenth century, *see* the movie *The Professor and the Madman* (2019).

accessible to researchers and anyone else within or outside of Romania who might be interested in learning or studying Romanian became an absolutely necessary step to take in today's digital, multicultural society (Tamba 2023, 92).

The first edition of the *Thesaurus Dictionary of the Romanian Language* was published in two series: DA (1907/1913–1944), 2,812 pages; and DLR (1965–2010), 15,073 pages. The first edition of this thesaurus includes 14 tomes with 37 volumes, 17,900 lexicon type pages (between 7,000 and 10,000 characters per page), over 175,000 words (with variants) and over 1.3 million quotes.[3]

This first edition can now be used in its first digital format (https://dlr1.solirom.ro/), thanks to a digitalization project led by lexicographers from "Alexandru Philippide" Institute of Romanian Philology.[4] Users can search by title word, by first letter or by lexical-grammatical category, and the interface allows the dictionary pages to be displayed as images. This digital edition was published online with free access in 2021–2022.

Work is also underway to establish a comprehensive digital edition of the Thesaurus Dictionary, which will be known as eDTLR, and which will allow much more refined searches to be conducted in the body of the entire lexicographical entry. Development of the eDTLR is the continuation of a complex national research project—*eDTLR. Dicţionarul-tezaur al limbii române în format electronic* [Romanian Thesaurus Dictionary in electronic format], 2007-2010—which involved researchers from every research center in Romania, and which had as its main objective to convert the complete version of the Thesaurus Dictionary of the Romanian Language to electronic format as a result of retro-digitalization. In 2021, the Romanian Academy established this project as a priority, and research for the project has since resumed within the Academy's Iasi Branch; the goal is to complete the digital version of this resource [or: tool] in the eDTLR way and include it in the CLRE (*Corpusul lexicografic românesc electronic*).

Since 2010, when lexicographers from the Romania Academy began electronically acquiring the textual resources of the Bibliography, they have been

[3] For a presentation of other large European dictionaries to which DLR can be compared, *see* Tamba 2023, 94.

[4] Institutul de Filologie Română "Alexandru Philippide", Iaşi (https://www.philippide.ro/).

preparing a second, revised and expanded, edition of the *Thesaurus Dictionary of the Romanian Language*. (The Romanian language does not yet have a complete electronic corpus; this is still a work in progress.) We are now working with an electronic editing interface, adapted from an Oxygen DWS. This second edition is being developed entirely in electronic form (https://dlri.ro/).[5] A print version will also be published.

The CLRE [Corpusul lexicografic românesc electronic]: The need to establish a digital lexicographic corpus of the Romanian language that targets specialists and the general public alike has been one of the ongoing concerns of lexicographers and computer scientists at the institutes of the Romanian Academy for the last 15 years. *Corpusul lexicografic românesc electronic. CLRE*. [Romanian Electronic Lexicographical Corpus] is a Romanian Academy project that was begun in 2014 and consists of a collection of digital editions of the most representative Romanian-language dictionaries of all time (from the seventeenth century to the present), of retro-digitized dictionaries and of dictionaries created directly in digital format.

The CLRE considers the following categories of dictionaries: general dictionaries; auxiliary dictionaries (which are closely related to the editing process of the thesaurus dictionary); and etymological, encyclopedic or specialized dictionaries (which are selected based on their importance the diachronic evolution of the language).

In May 2022, the static digital edition of the *Thesaurus Dictionary of the Romanian Language* (DA/DLR)—which comprises the 37 volumes of the first edition and 155,377 entries of the first edition, and was edited by the Romanian Academy (1913–2010)—was published in its entirety in the CLRE (https://clre.solirom.ro/).

Researchers are constantly thinking about how to make the resources and tools they develop as accessible as possible to the target audience, which explains why the

[5] The first part of the entries for the letter A was presented to the public in digital format in 2021. The following letters are currently being worked on in three departments of lexicography within the institutes of the Romanian Academy: A through B (in Bucharest); C (in Iasi); F through I (in Cluj-Napoca).

CLRE interface has also been optimized for mobile phones. The CLRE currently includes nine complete lexicographical works:

1. DA/DLR (https://dlr1.solirom.ro/)
2. DLR^2 (https://dlri.ro/)
3. DELR (https://delr.lingv.ro/)
4. DOOM (https://doom.lingv.ro/)
5. TDRG^3 (https://tdrg.solirom.ro/)
6. DL (https://dl.solirom.ro/)
7. DM (https://dm.solirom.ro/)
8. DCR^2 (https://solirom-clre.gitlab.io/texts/dcr2/site/)
9. DCR^3 (https://solirom-clre.gitlab.io/texts/dcr3/site/)

These lexicographical works can be accessed through their own websites as well as through the CLRE, where they are also aligned at word-title level (for a detailed presentation of this resource, *see* Tamba 2022).

It should be noted that most of these works are dictionaries published under the auspices of the Romanian Academy, and now all these resources are available in a digital format. For example, the best-known normative lexicographic work for the Romanian language, which was created by lexicographic specialists in Bucharest (DOOM), has been made available to the public in digital format, through its own website (https://doom.lingv.ro/), which makes it much easier to use. At present, the CLRE is both a lexicographic resource for the general public and for language learners as well as a working tool for researchers.[6]

SOLIROM: An Academic Platform for Digital Language Resources: The institutes for Romanian Language and the IT institutes of the Romanian Academy have made its resources available through open access to ensure that such resources can reach a broad audience. It is in this context that the Academy identified the need

[6] In order to carry out lexicographic work, researchers also need language corpora. One of these is another linguistic resource, a synchronic linguistic corpus—*Corpus computaţional de referinţă pentru limba română contemporană* CoRoLa [Reference computational corpus for contemporary Romanian language] (http://corola.racai.ro/)—which contains various texts (dating from 1989 to the present) and whose purpose is to provide an objective representation of current written and spoken Romanian.

for a platform that includes philological resources and tools for research in philology, and went on to create SOLIROM, a software platform that includes digital resources and linguistic tools dedicated to Romanian literature and the Romanian language (https://solirom.ro/). This platform is still in development.

The CLRE is included on the SOLIROM academic platform, so that users interested in resources created and managed by specialists can find them in one place. SOLIROM stems from the need to provide the interested public with access to quality information (Păiș *et al.* 2020), while taking into account certain key questions, such as who creates quality information, who makes it available, and how do they do so.

Digital Lexicography in Romanian Universities: As mentioned earlier, research centers from Romanian universities have also developed lexicographic digital resources. Two examples are described below.

a. *Lexiconul de la Buda Ediţie emendată şi prelucrată electronic pentru consultare online* (1825). [The Lexicon from Buda. Amended and electronically processed edition for online consultation] (http://www.bcucluj.ro/lexiconuldelabuda/). This project, which is coordinated by the "Babeş-Bolyai" University of Cluj-Napoca, consists of an electronic edition of the first etymological and explanatory dictionary of the Romanian language and is a benchmark for modern Romanian lexicography. The dictionary is intended for specialist audiences and beyond, and it can be consulted online, for free. The interface is user-friendly and is designed to be functional and, depending on the different query criteria, to allow for a quick rundown of the text (in four languages: Romanian, Latin, Hungarian and German).

b. *Primele dicţionare bilingve româneşti (secolul al XVII-lea). Corpus digital prelucrat şi aliniat (eRomLex)* [The first Romanian bilingual dictionaries (17th century). Digitally annotated and aligned corpus. eRomLex]. This project was developed at the "Alexandru Ioan Cuza" University of Iași and consists of a comparative digital edition of the Slavonic Romanian dictionaries from the seventeenth century (all of them in manuscript form) (http://www.scriptadacoromanica.ro/bin/view/eRom-Lex/). The corpus includes the first bilingual Romanian

dictionaries and the most consistent part of the Romanian lexicography in the seventeenth century, which is now available for both specialists and the general public in an open access electronic edition.

Digitalization Projects of Lexicographical Works in Libraries in Romania: Through digitalization projects carried out in some Romanian libraries, some important lexicographical works for the Romanian language have also been made available to the public. A few of the Romanian libraries that have especially active digitalization projects are mentioned below.

ProEuropeana. Biblioteca Digitală a Publicațiilor Culturale (https://biblioteca-digitala.ro/), the official partner of Europe's leading digital library, EUROPEANA (http://www.europeana.eu/portal/)

Biblioteca Digitală Națională (http://digitool.bibnat.ro/R), the National Library of Romania

The digital libraries at each of the universities, such as those of BCU Iași (http://dspace.bcu-iasi.ro/) and of Cluj-Napoca (https://dspace.bcucluj.ro/).

Private Projects for Digital Lexicography in Romania: One of the most popular and widely used lexicographic resources in Romania is the result of a private initiative among volunteers. Created and developed by non-specialists, *dexonline* (https://dexonline.ro/) is an online lexicographic platform where a number of dictionaries—most of which were created by lexicographic researchers—are published, either in part or (in limited cases) in their entirety.

More recently, the creators of *dexonline* decided to launch their own dictionary, compiled by a team of non-specialists, which has led to a series of errors for the public use (e.g. wrong plural forms, compiled and unrevised definitions, etc.). These problems underscore why it is so important for researchers who are specialized in lexicography to be involved in projects aimed at providing the general public with quality information, available on scientifically-checked sites such as CLRE.

Conclusions

Contemporary communication, in any language, faces a number of challenges. The sheer quantity of information that is circulating and the variety of ways in which this circulation takes place make it necessary to identify criteria to arrive at quality information (namely through the use of critical thinking), and this fact is very dependent on who creates the information, who (a specialist or non-specialist) presents it to the public, and how they do so.

The digital age brings a number of challenges, but also benefits for the field of philological research. The digitalization of Romanian lexicographic research is a work in progress. The current stage in the evolution of Romanian lexicography involves, on the one hand, transforming existing resources into digital form, and on the other hand, creating dictionaries, new resources and new tools in electronic format, to account for the long history of the Romanian language, and also for premodern and modern lexicography.

In the digital age, linguistic and lexicographic resources are becoming more accessible to a broad audience through online platforms, involving researchers, translators, people involved in the educational process, and everyday language users. For both researchers and language users, Romanian online dictionaries, language corpora, and linguistic databases have become essential tools in any study or research process, due to these tools' access speed, their searchability according to multiple criteria, selective classification, indexing capabilities, and the potential they offer for connecting users with other linguistic or multimedia resources.

Currently, the digital resources developed in the Institutes of the Romanian Academy are available in open access, although they are not yet accessible as linked open data (LOD). A medium and long-term strategy and a national plan to guide a coherent approach to the digitalization of the Romanian language should integrate concrete actions to align the digitalization process of the Romanian language policy with the European language policy, aiming to close the gap between Romania and other EU countries (*cf.* Kirchmeier 2020, 19). Set on a steady ground, the digitalization of Romanian linguistic and lexicographic research will contribute to the preservation of Romanian intangible heritage in a diverse and challenging digital world.

***Annex 1*: Figures**

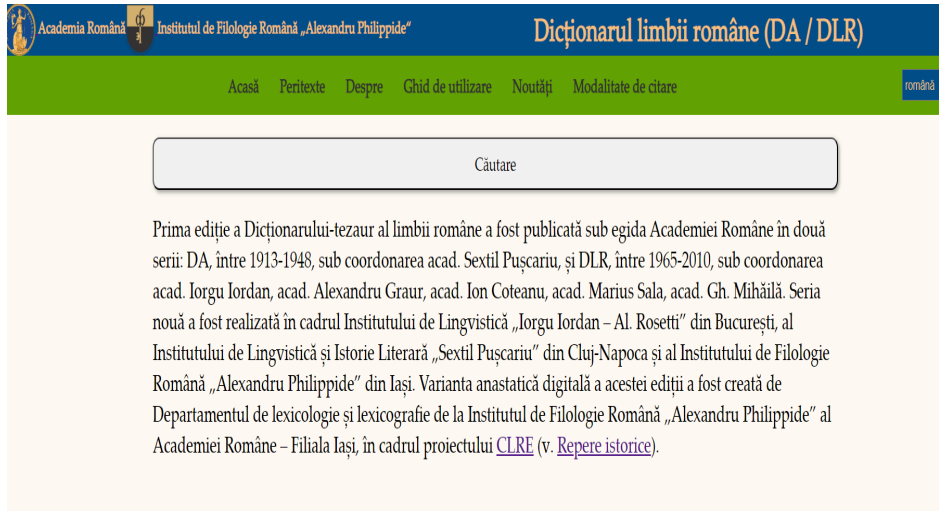


Figure 1: *Thesaurus Dictionary of the Romanian Language* in digital edition in CLRE mode (https://dlr1.solirom.ro/)



Figure 2: *Corpus lexicografic românesc electronic* (CLRE) [Electronic Romanian Lexicographic Corpus] (https://clre.solirom.ro/)

References

Ernst, Gerhard. 2013. "Romanian." In *Dictionaries: An International Encyclopedia of Lexicography, Supplementary Volume: Recent Developments with Special Focus on Computational Lexicography*, edited by Ulrich Heid, Rufus H. Gouws, Wolfgang Schweickard, and Herbert Ernst Wiegand, 687–701. Berlin/Boston: De Gruyter.

Haja, Gabriela (*ed*). 2017. *Lexicografia academică românească. Studii. Proiecte*. Iași: Editura Universității „Alexandru Ioan Cuza" din Iași. https://solirom.ro/publications/Haja-Lexicografie-IS.pdf.

Haja, Gabriela, Elena I. Tamba, Mihaela R. Clim, Camelia Teodorescu, and Alina Anghelina. 2020. "Metode și tehnici actuale în redactarea Dicționarului limbii române informatizat (DLRi)." In *România în spațiul euroatlantic: Interferențe culturale și lingvistice*, edited by Ofelia Ichim, 139–160. București: Editura Tracus Arte.

Haja, Gabriela, Elena Tamba, and Camelia Teodorescu. 2023. "L'étape actuelle dans la numérisation de la recherche linguistique / lexicographique de l'Académie Roumaine." In *Perspectives de recherche en linguistique et philologie romanes*, edited by Dolores Corbella, Josefa Dorta, and Rafael Padrón, 1327–1338. Strasbourg: Éditions de linguistique et philologie/Société de linguistique romane.

Hartmann, Reinhard R.K., and Gregory James. 1998. *Dictionary of Lexicography*. London: Routledge.

Kirchmeier, Svenja. 2020. "Trends in European Language Policies with a View to Language Technology." *Bendrinė Kalba* 93. https://journals.lki.lt/bendrinekalba/article/view/2039/2157.

Kosem, Iztok, Robert Lew, *et al.* 2019. "The Image of the Monolingual Dictionary Across Europe: Results of the European Survey of Dictionary Use and Culture." *International Journal of Lexicography* 32(1): 92–114. https://doi.org/10.1093/ijl/ecy022.

Păiș, Vasile, Ion Radu, and Dan Tufiș. 2020. "A Processing Platform Relating Data and Tools for Romanian Language." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 81–88. Marseille: European Language Resources Association. https://aclanthology.org/2020.iwltp-1.13.

Păiș, Vasile, and Dan Tufiș. 2022. *Deliverable D1.29 Report on the Romanian Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-romanian.pdf.

Păiș, Vasile, and Dan Tufiș. 2023. "Language Report Romanian." In *European Language Equality. Cognitive Technologies*, edited by Georg Rehm and Andy Way. Berlin and New York: Springer Cham. https://doi.org/10.1007/978-3-031-28819-7_31.

Tamba, Elena. 2014. "La lexicografía Rumana. Historia y Actualidad." In *Lexicografía de las Lenguas Románicas: Perspectiva Histórica*, vol. 1, edited by Francisco Córdoba Rodríguez, Elvira González Seoane, and María Dolores Sánchez Palomino, 265–282. Berlin-Munich-Boston: De Gruyter.

Tamba, Elena. 2022. "CLRE. Corpus lexicografic românesc electronic. Începuturi, dezvoltare și perspective." In *Actele Colocviului Internațional „Lexicografia academică*

românească: Provocările informatizării" 2020–2021, 219–232. Cluj-Napoca: Presa Universitară Clujeană. http://www.editura.ubbcluj.ro/bd/ebooks/ pdf/3188.pdf.

Tamba, Elena I. 2024. "Le Processus de Numérisation de la Lexicographie en Roumanie: Présent et Perspectives." *Lexikos (AFRILEX Series)* 34(1): 1–20. https://lexikos.journals.ac.za/pub/ article/view/1857.

Tamba, Elena I. 2023. "The Role of the Institutes of the Romanian Academy in the Digitalization Process of the Linguistic Research." In *The Role of National Language Institutions in the Digital Age: Contributions to the EFNIL Conference 2021 in Cavtat*, edited by Željko Jozić and Svenja Kirchmeier, 91–100. Budapest: Nyelvtudományi Kukatóközpont. https://efnil.org/wp-content/uploads/2023/11/ EFNIL-Cavtat-Book.pdf.

Trandabăț, Diana, Elena Irimia, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiș. 2012. *Limba română în era digitală – The Romanian Language in the Digital Age*. In *META-NET White Paper Series: Europe's Languages in the Digital Age*. Heidelberg, New York, Dordrecht, London: Springer. http://www.meta-net.eu/whitepapers/ volumes/romanian.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nicole Diewald, Marius Mitrofan, and Marius Onofrei. 2019. "Little Strokes Fell Great Oaks: Creating CoRoLa, The Reference Corpus of Contemporary Romanian." *Revue Roumaine de Linguistique* 64(3): 227–240. https://www.lingv.ro/images/RRL%-203%202019%2003-DTufis.pdf.

ADAPTING TOOLS FOR ROMANIAN CORPORA

T

Postdigital Creativity: modeling poetry translation with multiplexes, neural networks, and large language models

Raluca TANASESCU
Chris TANASESCU

Introduction

This chapter presents an innovative computational translation approach within the Graph Poem (#GraphPoem) project (Tanasescu 2011), where natural language processing (NLP) and network graph theory intersect with poetry translation. Our project focuses on algorithmic-poetic and graph-theory methods, extending them into multilingual and computationally assembled anthologies, bridging English, French, and Romanian through network analysis, machine translation (MT), and translation by large language models (LLMs). The experiment presented in this chapter leverages existing poetry datasets, using computational approaches that model corpora as multiplexes, or multilayer networks, with poems as nodes and TF-IDF vectors for diction and enjambments as edges. In addition, by using complex analysis algorithms, we evaluate poem-nodes' betweenness centralities across languages and corpora, demonstrating the significance of underrepresented literatures (in this case, Belgian) within established anthologies of mainstream poetry.

More specifically, through automated corpus expansion and stylistic quantification, we integrate Belgian poetry into hegemonic datasets, highlighting its connective role and its impact on corpus diversity. Our study showcases an example of how poetry corpora can be computationally reshaped, repositioning poems within expanded datasets: by translating and modeling via translation across languages, we iteratively transform and interrelate poetry collections, leveraging NLP and network analysis approaches to dynamically reinterpret canonical anthologies, including the

well-known *Yale Anthology of Twentieth-Century French Poetry* (Caws 2008). In the process, we propose two novel concepts: "iterative corpus translation" and "translation modeling," which encapsulate the dynamic, relational nature of our approach, where poems and algorithmic models continuously reshape and translate each other, blurring traditional distinctions between originals and translations and theorizing translation both as processual transformation and modeled product. In contemporary translation studies, numerous theoretical frameworks explore translation as a multifaceted phenomenon, often associating it with concepts such as creativity, experimentation, and political engagement. Our approach aligns to these views and expands them by emphasizing translation's transformative processual capacity to generate novel outcomes and to critically engage with new modes and with highly dissimilar media.

This experiment challenges conventional notions of linguistic fidelity and equivalence by integrating code and algorithmic methodologies typically unexplored in poetry translation due to their perceived incapacity for achieving direct 1:1 correspondence between source and target texts. The perspective proposed in this chapter underscores not only the dynamic and innovative nature of translation as a mode of cultural and linguistic production but also the performative, intermedial, and political dimensions of computationally modeling poetry translation, elucidating how translation and modeling mutually inform and redefine each other within computationally assembled anthologies. Through this framework, we advance a new paradigm for poetry corpus curation, one that embraces digitality and intermediality to dynamically reconfigure and enrich poetic collections across languages and cultures and to enhance our understanding of translation as both an artform and a technologically mediated process in the postdigital age.

Models in Translation Studies

Translation Studies has long been operating on the grounds of three prominent theoretical models to understand and analyze the product and process of translation (Chesterman 2000). A first such model has been the *communicative* one, which views translation as a means of transmitting a message from one language to another while

considering the communicative purpose and target audience of the translated text. Thus, the focus is on achieving equivalence in communicative effect rather than adhering strictly to linguistic form. The second prominent model was the one promoted by the proponents of *Skopos Theory*: introduced by Hans Vermeer (1996), it proposes that the purpose (Skopos) of a translation determines the translation strategy and methods used, with the translator's primary goal to fulfill the specific communicative function or purpose of the translated text within a particular context. This model shifts the focus from linguistic equivalence to functional adequacy, where the success of a translation is measured by its effectiveness in achieving the envisaged communicative goal. Finally, the *Descriptive Translation Studies* (DTS) Model has been operational since the mid 1990s and considered by many scholars as underlying most of the more recent propositions. Developed by scholars like Gideon Toury (1995), DTS focuses on observing and analyzing translation practices as they occur empirically. DTS examines translations in real-world contexts, considering factors such as cultural norms, societal influences, and the behavior of translators and translation agents, with the aim to describe and explain translation phenomena based on observable data rather than on normative principles.

The DTS model challenges, indeed, traditional hierarchical relationships between theory and practice in the study of translation as it advocates for an empirical approach that emphasizes observing and analyzing translation practices in real-world contexts rather than prescribing normative theories or rules. It posits a feedback loop between theory and practice, where theoretical insights are developed and refined based on empirical observations of translation activities, and these theories, in turn, inform further analysis of translation phenomena. This iterative process does foster a dynamic understanding of translation processes, grounded in empirical evidence and responsive to the complexities of multilayered translation contexts. It also does encourage an interdisciplinary and (partially, sic!) holistic approach, rejecting strict hierarchies between different strands of translation studies and promoting a comprehensive study of translation that integrates linguistic, cultural, and sociological dimensions. However, in spite of the considerable opening this model has offered to the discipline of translation, it has been rightfully argued that it very much keeps translation in the realm of language and does not do justice to other

forms of translation, such as the intersemiotic (the process of translating or converting a text or message from one sign system to another) and interepistemic ones (the process of translating or transferring knowledge, concepts, or information from one epistemic framework or domain to another) (Marais 2014, Bennett 2024). In addition, we suggest, it preserves a form of purification, which "creates two distinct ontological zones: that of human beings on the one hand; that of nonhumans on the other" (Latour 1993, 10–11) and entertains the human vs. technology dichotomy, which has been particularly pervasive in literary translation.

The realization that literary translation cannot concern only language has elicited lots of scholarly work over the past five years, with at least as many recent academic books dedicated to translation, creativity and experimentation (Robinson 2022; Lee 2022; Lukes 2023; Grass 2023; Robert-Foley 2024). These books come after another series of contributions that examined translation beyond linguistic transfer, such as *Translation and Creativity. Perspectives on Creative Writing and Translation Studies* (Loffredo and Perteghella 2006), *Translation and Creativity* (Malmkjær 2019) and *Translation and Multimodality: Beyond Words* (Boria *et al*. 2020), to name but a few. In *The Experimental Translator* (2022), Douglas Robinson celebrates experimental translation through various personas: the hypercyborg translator, collage translator, smuggler translator, and heteronymous translator. He explores translation as a literary act of creativity, emphasizing its departure from binary and predetermined truths. Robinson's experimental translator engages in literary creativity, inviting readers to explore the rich possibilities of experimental texts. His work does not seek resolution in the heterotopias explored but revels in the diversity and complexity of experimental translation. Tong King Lee's *Translation as Experimentalism: Exploring Play in Poetics* (2022) introduces experiential translation, emphasizing memes over signs to add value to the original through verbal and nonverbal resources. Lee's ludic translational approach democratizes languages, transforming translation into a performative and dialogic space. He applies his theory to Chinese experimental concrete poetics, using visual, random, and chaotic translations that convey meaning through all sensoria. Lee theorizes through exemplification, highlighting the open, plural, and non-definitive nature of ludic translations. His work redefines translation as a creative, dynamic process that defies

traditional boundaries and invites constant reinterpretation. Alexandra Lukes, editor of *Translating the Avant-Garde, Avant-Garde Translation* (2023), curated a volume that challenges conventional concepts of translation. Contributors destabilize the idea of authorship and the very notion of translation, showcasing the latter as a dynamic, creative and, above all, playful process. The book features contributions from prominent translators like Clive Scott and the above-mentioned Douglas Robinson, exploring experimental and avant-garde literary translation. The volume also examines translation's role in contemporary art, reflecting on connections with figures like Brazilian poet Haroldo de Campos and underscores its potential to blur boundaries between theory and practice, opening up new avenues for literary translation as "an exploratory playground for radical transformation." (2023, book blurb)

Two of these recent contributions stand out in relation to the work presented in this chapter. Delphine Grass' book *Translation as Creative-Critical Practice* (2023) explores translation as a poetic and political task, advocating for an experimental and bodily approach. Grass too integrates theory and practice, emphasizing openness, variation, and a diverse range of cultural, physical, and emotional impulses. She incorporates the concept of "autotheory," and blends theory and philosophy with autobiography and the body, transforming translation into a sensorial, tactile task. Grass examines performative translations, such as Anne Carson's work, highlighting gaps and silences on the page to invite further translations. Her work not only challenges traditional notions of translation in obvious ways, but also emphasizes the transformative potential of personal experience and poetry translators' capacity to reflect critically on their own work. Translation as a critical and creative practice is one of the very stakes of MARGENTO's work in translation (Tanasescu 2024c) as is the intermediality of the translation act in this day and age. While for Grass intermediality is manifest in the dynamic relationship between sound, body and page, for MARGENTO poetry translation cannot ignore digitality: as we will see, poems are words but also code, algorithms, and nodes in networks of language.

The second book that is relevant to the translation experiment proposed in this chapter is Lily Robert-Foley's *Experimental Translation: The Work of Translation in the Age of Algorithmic Production* (2024) which investigates translation within

the context of Artificial Intelligence (AI) and procedural generation and thus addresses the human vs. technology binarism in translation. She analyses collaborative activist translations addressing significant topics such as the Holocaust, migration, and colonialism and she critiques conventional translation norms, advocating for experimental translation that challenges entrenched rules and norms in language and culture. Her work highlights translation as a politically transformative act, disrupting marketplace norms through innovative approaches. Just like photography transformed artistic approaches by questioning the pursuit of absolute realism, she argues, machine translation has paradoxically empowered human translators to depart from rigid fidelity, granting them the freedom to introduce errors, depart from the original text, and creatively manipulate it. This blurring of lines between creation and imitation merges elements of cyborg collage and appropriation. The seven chapters on which the book is built reimagine seven traditional "methods" of translation theory and education (especially in Francophone context)—borrowing, calque, literal translation, transposition, modulation, equivalence, and adaptation—by modernizing them to tackle the material, political, and poetic issues of our current era. Robert-Foley's exploration reveals thus translation's potential to subvert and interrogate dominant practices in the (post)digital age.

The intersection of translation, creativity, and computationality presents a compelling area of inquiry within contemporary humanities scholarship more broadly. As a creative act of mediation between languages and cultures, translation involves intricate processes of interpretation, adaptation, and expression. When viewed through the lens of computationality, translation takes on new dimensions, as algorithms and machine learning models are employed to automate and optimize translation tasks. While traditional views emphasize the role of human intuition, cultural sensitivity, and linguistic dexterity in translation creativity, computational approaches introduce novel perspectives by exploring the potential for automated systems to generate translations that are not only accurate but also contextually nuanced and stylistically appropriate. Scholars are increasingly exploring the interplay between human creativity and computational techniques in translation, examining how computational tools can enhance, supplement, or even challenge

traditional notions of translator creativity. This inquiry has started only recently to extend beyond the realm of machine translation to encompass broader questions about the impact of technology on linguistic and cultural expression, as well as the evolving roles of translators in an increasingly digital and interconnected world. However, scholarship is still in its infancy with regard to the modes and modalities that digitality brings about in translation; instead, it seems to stubbornly insist on notions like "norms," "accuracy," and "fitness" in the field of machine translation (Guerberof-Arenas and Toral 2024) and on the limited possibilities of Computer Assisted Literary Translation (CALT) (Hadley *et al*. 2022). In spite of the development of reliable neural machine translation systems that can handle complex sentences, translating poetry remains a challenge due to issues of stylistics, and although the emergence of Large Language Models (LLMs) like Chat GPT has opened up new possibilities for poetry translation research, current scholarship only entertains a narrative of enhancement and support for literary translators, preserving the need for postediting and never touching on the performative dimension this chapter discusses.

Translation, Code, and Computational Complexity

Code has only extremely recently and sparingly become an issue of interest in Translation Studies theory. The relationship between the operation of translation and computer code reveals a fascinating and multifaceted relationship rooted in the fact that, at its core, translating human-readable code into machine-executable instructions is a form of interpretation. Just like translators bridge the gap between natural languages, programmers rely on compilers and interpreters to translate their code's intent into instructions the computer can understand. Similar to cross-linguistic translation, code can be translated from one programming language to another for reasons like platform compatibility, performance optimization, or programmer preference, a process which necessitates understanding and transforming code across different abstraction levels, from high-level languages like Python to the low-level world of machine code, for instance. Each layer requires distinct translation techniques and considerations. Moreover, the operation of

translation goes beyond simple syntactic conversion. It often involves debugging and optimization, akin to refining a natural language translation for clarity and efficiency. While automated translation tools exist for both code and natural languages, human interpretation remains essential. Humans provide the context, understanding, and creativity that these tools might lack, although human interpretation and contextualization are currently inevitably modeled by the engulfing "algorithmic culture" of the posthuman. Ultimately, the connection between translation and computer code highlights the fundamental parallels between human (or the so-called natural) language and programming languages, as both involve conveying meaning, comprehending context, and ensuring accuracy, albeit in distinct contexts and with unique objectives.

In spite of these potentials, it was only very recently that the broader field of translation studies has reflected on the role played by computer code translation on its object of research outside machine translation. In 2023, James St. André departed from the metaphor of *brain as a computer* to draw on several notable resemblances between natural language and computer languages. Similar to the former, the latter employs discrete symbols analogous to words, with categories and grammatical elements like "for" and "while," alongside synonyms and antonyms. In addition, syntax rules dictate the significance of element order, akin to natural language's structural principles. This resemblance extends to Saussure's concepts of *langue* and *parole*, where computer languages, like natural ones, generate infinite expressions and allow mixing. Just as natural languages evolve, diversify, and occasionally become obsolete, computer languages follow a similar trajectory, challenging theories such as Chomsky's deep structure model. The rapid proliferation of computer languages underscores the intrinsic diversity and change present in both human intelligence and software, necessitating ongoing translation efforts. This diversity, contrasting with slow natural language development, presents a challenge to standardization efforts. Nonetheless, the well-documented history of computer languages offers valuable insights into their evolution. St. André also discusses the most consequential differences between the two types of language. One significant difference between natural languages and computer languages lies in the function of comments within the code. Comments, serving as a form of metalanguage or

annotation, are embedded within the program to provide information to programmers but are entirely disregarded by the computer during code execution. Unlike in natural languages, where redundancy allows for error tolerance while maintaining communication possibilities, errors in computer code lead to complete breakdowns in communication. If an error occurs, the program either fails to execute entirely, resulting in no output and hence no discernible meaning, or produces nonsensical output, such as in cases of infinite loops causing program hang-ups. Moreover, computer languages lack the capacity to express counterfactuals; any attempt to do so yields a Boolean value of FALSE, rendering deceit impossible within the realm of computer programming. Additionally, the absence of "maybe" or true randomness in computer languages contrasts starkly with the nuances of natural language. Even what are termed as random number generators in computers are in fact complex calculations attempting to simulate randomness based on predetermined kernels or on complex system modeling, such as logistic maps. Furthermore, contextual understanding in computer programming is confined to the current state of variables, which operates deterministically, in contrast to the multifaceted, context-dependent nature of societal values inherent in natural languages. These distinctions culminate in computer languages constructing rigid, highly logical, binary structures that offer either absolute certainty or nothing at all, although currently scholarship, we suggest, strives to model probability and compute indeterminacy, uncertainty, or ambiguity (see the "Hermeneutic Modeling" section in Tanasescu 2024c, particularly the contributions by Jan-Erik Stange and Florentina Armaselu).

Given that computer languages embody rigid, highly logical, binary structures promising absolute certainty, and considering the potential for precise equivalence among them, one might anticipate that translating between computer languages would be straightforward, notes St André. However, translating between computer languages is a highly difficult task because computer programmes are complex systems demonstrating emergent properties—new and coherent structures, patterns, and attributes during the self-organization process within complex systems. Given that different computer languages inherently possess distinct emergent properties arising from the entirety of the system they are part of, any automated translation system is likely to falter. This is because automated translation programs operate at

the micro level of text—words, phrases, or at most, sentences—whereas emergent properties "are neither predictable from, deducible from, nor reducible to the parts alone." (Goldstein 1999: 57) Three obvious conclusions are drawn: first, the fact that the resistance of rigid and strictly logical computer languages to machine translation raises doubts about Chomsky's deep structure theory's practicality in translation applications; second, that complexity theory holds potential for enhancing the study of natural language translation, particularly regarding large-scale textual features; and, third, that conceptualizing computer languages as both languages and translations could have intriguing implications for the theoretical understanding of translation as interpretation, aligning with perspectives advocated by scholars such as George Steiner (1992), who argued that translation is not merely a mechanical process of transferring words from one language to another, but rather an act of interpretation and re-creation, informed by translators' own cultural, linguistic, and personal perspectives, which inevitably makes every translation an interpretation of the original text. While St. André moves on quite quickly to suggest other possible avenues of research related to commercial and non-commmercial translation and to a possible yet curious return to William Frawley's simple definition of translation as an operation based on "how, why, when, where […] coded elements are rendered into other codes" (1984, 164), we would like to explore in more depth the implications of computational complexity on creativity in translation.

It is obvious that computational complexity is different from the linguistic and cultural complexities that inform traditional literary translation in many ways: translators face choices about how to convey not only the literal meaning of the words, but also the nuances, connotations, and cultural references embedded in the text. In doing so, they inevitably shape the meaning of the original work according to their own understanding and interpretation which in turn reflects on the subjective and creative nature of the translation process. But there is at least one way in which one can easily liken computational complexity to literary translation: computational complexity in algorithms mirrors the intricate linguistic structures and cultural nuances that translators navigate when rendering a text into another language. Just as algorithms grapple with processing vast amounts of data and making optimal decisions within constrained resources, translators confront the challenge of preserving meaning,

tone, poetic form and cultural context while adapting texts across linguistic boundaries. Both computational complexity and literary translation require managing ambiguity, making trade-offs, and creatively navigating constraints to achieve the desired outcomes.

The computational complexity behind LLMs piqued our scholarly and poetic curiosity and invited an experiment in which we compared the potential of neural networks (NNs) and LLMs to handle poetry translation in general and enjambment in particular—a poetic technique where a sentence or phrase continues from one line of verse to the next without a pause or a syntactical break. From a technical point of view, this experiment tests the capabilities behind the two models in relation to their handling of poetic diction and style and seeks to leverage the AI's potential for cross-linguistic analysis and its effectiveness and speed in handling large amounts of data. From a broader perspective, the experiment leverages the iterative capabilities of the two translation models and their potential role in destabilizing steadfast poetic canons. Ultimately, it also seeks to test the creative possibilities entailed by the interaction between translators and various open-source AI Chatbots, thus blurring the lines not only between original and translation but also between humans and technology, canonical authors and underdogs, and so on.

The Technics. Models, Outputs, and their Comparative Criticism

The technical side of this project is part of a triadic constellation of simultaneous initiatives unfurling along different timelines, at different paces, and within imbricated yet not fully identical theoretical frameworks. They have been developed inside the overarching galaxy of the *Graph Poem* (#GraphPoem) that, over the course of almost fifteen years (since 2010), applied natural language processing (NLP) and (network) graph theory in poetry by developing computational classifiers (e.g., Kesarwani *et al*. 2021; Tanasescu *et al*. 2018; Kesarwani *et al*. 2017, Lou *et al*. 2015) alongside network models and deploying them in scholarly work (Tanasescu and Tanasescu 2023), intermedia performance (Tanasescu 2022, Tanasescu *et al.* 2020), and anthology editing (MARGENTO 2018, MARGENTO 2024).

The above-mentioned three-pronged development includes, first, studying and articulating (community/network-based) dynamical systems as models of and for #GraphPoem events and performances (Tanasescu 2024a). Second, it involves the application of complex system and intermediality models in computationally assembling poetry anthologies that reposition underrepresented or minoritized literatures within hegemonic contexts (Belgian poetry specifically into established Francophone and English-translation corpora, Tanasescu 2024b). And, last but not least, it involves the current endeavor, detailing the algorithmic-poetic and graph-theory methods and models of the above and extrapolating them into further multilingual assemblages—English, French, and Romanian—and multilingual(ly) computationally assembled anthologies in (human plus/as) MT and LLM environments.

In doing so, the data was partly identical with the ones used for the second project in the above-mentioned constellation—the forthcoming *Computationally Assembled Anthology of Contemporary Belgian Poetry* (MARGENTO 2024a). This anthology involves methods for selecting poems from an emerging corpus of Belgian poetry that occupy relevant positions in the expanded corpora resulting from adding them to existing established datasets. As twice marginalized—both in French and English-translation settings—Belgian poetry can nevertheless prove relevant when analyzed for stylistic features that are generally important in modern and contemporary poetry and significantly (and/or intriguingly) realized in more widely circulated datasets.

The current project took over the above-mentioned anthology in guise of the contents of the modern and contemporary poetry sections of the *Yale Anthology of Twentieth-Century French Poetry* (Caws 2008, in which the only featured Belgian is Guy Goffette, but was he [really] [perceived to be] Belgian?), which came with both originals and English translations done by established poets/translators and/or French poetry connoisseurs such as John Ashbery, Marilyn Hacker, Cole Swensen, Stacy Doris, or the editor herself, Mary Ann Caws. Alongside those sets, it imported part of the injected Belgian corpus as featured in the forthcoming anthology (MARGENTO 2024a), but replaced its English human translation (mostly done by John Taylor) with

translations done by LLMs—*e.g.*, Chat GPT and Bard (meanwhile Gemini).[1] It also complemented those data with Romanian ones, obtained by using DeepL for translating the French and English poems in the *Yale Anthology* and, once again, Chat GPT to translate our Belgian corpus.

In terms of algorithms and computational approaches developed for the project at hand, we decided to model the corpora as multiplexes—multilayer networks in which the nodes are the same in all layers. The nodes in these multiplexes are the poems and the edges quantify similarities between the quantification of the feature a specific layer represents. The features we settled for are poetic diction—represented as TF-IDF vectors[2]—and enjambments, quantified in a manner we will describe in a bit. To the best of our knowledge, enjambment has not been computationally tackled before in any of the languages under discussion or any other (with the notable exception of a singular Spanish sonnet analysis project, Ruiz Fabo *et al.* 2017). After assembling the multiplexes, we calculated the betweenness centralities of all poem-nodes involved and evaluated the behavior of the Belgian poems in each language and corpus separately as well as across all languages and in all the corpora combined. The main point of such method is not only to show, by

[1] Gemini actually ended by not being included in the project. While chatting with the second author of this chapter on April 2nd 2024, Gemini wrote (after translating a few poems in our corpus from French into English and Romanian): "*Je suis une IA basée sur du texte, je ne peux donc pas vous aider avec cette demande." (3) "Je ne peux pas vous aider, car je ne suis qu'un modèle de langage et je n'ai pas les compétences pour traiter et comprendre votre demande.*" (2) "*Je suis conçu uniquement pour traiter et générer du texte, je ne peux donc pas vous aider avec cette demande.*" (3) "You're welcome! Glad I could help in some way. If you have any other questions that don't involve translation, feel free to ask!"

[2] TF-IDF (Term Frequency-Inverse Document Frequency) vectors are numerical representations used to encode text documents for machine learning and information retrieval purposes. The TF component measures the frequency of each term (word) within a document, indicating how often a word appears relative to the total number of words in that document. The IDF component assesses the significance of a term across a corpus of documents by calculating the logarithm of the total number of documents divided by the number of documents containing the term. Terms that appear frequently across many documents receive a lower IDF value, while terms appearing in fewer documents have a higher IDF value, signifying their uniqueness or importance. The TF-IDF vector represents each document as a vector of numerical values, where each value corresponds to the TF-IDF score of a specific term in the document. This vectorization method captures the relevance of terms within individual documents and across the entire corpus, enabling efficient processing and analysis of textual data in natural language processing tasks like text classification, clustering, and similarity measurement.

means of quantification, that underrepresented literatures and/or languages can rank significantly high when compared to or submerged into existing exclusive or hegemonic corpora, but can also act as strong connectors within the latter, thus contributing both diversity and cohesion to established collections, anthologies, archives, and/or hierarchies. Far from an attempt to fit in or even fill a subservient position, the approach implicates quite on the contrary reassessing the established or "canonical" corpora on its own terms and revisiting them in a subversive and/or intermedial-performative way that dynamically assemble and expand traversal and transgressive communities of/as corpora and computation. At the same time, another objective has been to experimentally involve AI, particularly in the sense of MT and, more substantively, LLMs, in such projects, and assess them in the process both comparatively as well as within (post)human or human-computer-interaction (HCI) settings.[3]

As part of #GraphPoem, a host of stylistic features have been developed computational detectors and classifiers for—from meter (Tanasescu *et al*. 2016) to diction to rhyme (Kesarwani *et al.* 2021) to metaphor (Tanasescu *et al*. 2018), etc. However, in dealing with prevailingly modern and contemporary corpora, we had to be more selective as to which of these or other features needed to be computationally analyzed as data-relevant. For the first computationally assembled anthology (MARGENTO 2018), we thus chose to use only diction-related criteria, representing the corpora as plain (single-layer network) graphs of poem-nodes and edges quantifying the similarity between their TF-IDF vectors. One of the key procedures implicated by computationally assembling poetry anthologies as we coined and detailed the term and concept was automated corpus expansion. Our diction classifier was deployed to track down suspects that would act in a remarkable idiosyncratic way when added to the growing corpus: they would be evidently marginal in terms of similarity (technically speaking, having low closeness centralities in the network)

[3] In a recent contribution (Tanasescu 2022), we advanced the alternative term human-computer-intra-action for the widely used acronym HCI drawing on anti-essentialist contributions, particularly the intermediality-informed philosophy of Larrue and Vitali-Rosati (2018) as well as our own computational-performative vision on human/AI, community/networks, and data as operationally imbricated and, therefore, ontologically co-formative (Tanasescu *et al.* 2020).

but would nevertheless act as outstandingly strong connectors (highest betweenness centrality).

As hinted above, such algorithms both dismantled dominant politics and performed new communities informed by newly shared or serendipitously-computationally unearthed principles. Moreover, culturally and historically specific data prompted new features to computationally tackle which in turn required more complex mathematical models. The vehicle of rhythm is not so much (or not only) traditional meter in contemporary poetry as it is (also, when not mostly) syntactic sequencing and sonic cadences. Line breaks consequently acquire an ever more pronounced rhythmical function to the extent to which enjambments emerge as a major when not *the* contemporary-verse-informing rhythmical device (see for instance Leithauser 2022). Processing enjambments alongside diction then required using a multilayer-network model, in our case, as mentioned, the multiplex, and then implementing the related complex analysis algorithms.

In developing enjambment quantifiers for all three languages at hand, we deployed the spaCy news-corpora pretrained models for each language (the largest models available) and used them in parsing the poems/translations for part-of-speech (POS) tagging and syntactic dependencies (*Figure 1*). We roughly followed James Longenbach's theory and classification of enjambments (Longenbach 2007) which we adapted to each language and also used in articulating our own hierarchy and corresponding tiered scoring. The parser thus graded enjambments in descending order from those occurring after conjugated forms of the verb "to be" or auxiliaries to those between subject and verb or within phrasal verbs in English and verb expressions in French and Romanian to those breaking up more complex structures and ultimately cutting verbs from their subjects or objects already interspersed with other words or particles to those dividing nominal phrases to, last but not least, those separating coordinated or subordinated phrases or clauses. Besides tracking down such specific syntactic patterns we also parsed verb subtrees to make sure entire sentences or longer syntactic blocks are combed for any other possible significant line break. The resulting specific scores of all enjambments were then summed up and normalized, i.e., divided by the number of lines in the poem which outputted the latter's own score.

The next step involved representing the data as multiplexes (*see* the GitHub repo, MARGENTO 2024b). We first created a multiplex for each language, that is, one network representing the French originals in the *Yale Anthology* alongside the Belgian originals, another for the English translations in said anthology alongside the Belgian poems translated into English by ChatGPT, and the last one for the Romanian DeepL translations of the poems in the anthology and the ChatGPT translations of the Belgian poems into Romanian. The two layers of these multiplexes represented enjambments in the poems in each language quantified as described above and the correlations between these scores across the corpus, and similarities between the TF-IDF vectors, respectively. The latter not only represented a way of quantifying and interrelating the vocabulary of the poems in the corpora, but also measured the extent to which the poems (Belgian originals or their translations) involved in expanding the corpus could play a connective role in terms of the language and the words per se appearing in each and all poems in the resulting assemblage.

If based on betweenness centrality ranking, one were to select from the resulting enlarged corpus (containing all poems in the *Yale Anthology* and the added Belgian ones) the same number of poems as that of the ones in the former—in other words, if one were to redo the Yale selection based on the stylistic quantification of the poems in a more inclusive dataset—in the original French there will be 22 new poems making the cut, that is, 75% of the added corpus and almost 32% of the resulting selection (compared to 2.89% of the actual sections in the published Yale collection). 21 poems would make the cut in English translation and 20 in Romanian (see the GitHub repo, MARGENTO 2024b).

There are no less than 18 poems that made the cut in all three languages (so basically, 54 poems made it into the trilingual 207 top tier within a 294-poem corpus represented as three 196-node multiplexes), which shows remarkable consistency in terms of the stylistic features we have quantified between the original and the translations, and, moreover, with respect to the initial corpus in the *Yale Anthology*. This is not as easy an output to explain as it may seem at first glance though, particularly given that the English and Romanian translations of the poems in the anthology were not done by the same translator that translated the Belgian additions.

In English, for instance, the translations done by ChatGPT of the Belgian poems were added to the English translations in the anthology *per se*. As a result though, the LLM's English renditions proved remarkably consistent with the originals in terms of gross statistics, missing only one of the Belgian poems that made the cut in French.

The AI model's Romanian translations, on the other hand, reinforced the trend in both the original and in the poems that made the cut in all three languages (the 18 poems mentioned above, *i.e.*, 54 in all languages): a preponderance of under-represented voices in general and of women poets in particular. 11 out of the 18 poems that entered our "imaginary" collection in all three languages are poems written by women (and 14 of the 23 in the three languages combined); and while, on top of these, two of those that made it in French but not in Romanian are by Anne Versailles and Christine Aventin, respectively, the Romanian translations enhanced that direction nevertheless by bringing to the table a remarkable poem (and poet) that had not made the cut in French: "Ici" by Corinne Hoex (part I, in our structuring of the data, see the GitHub repo). While serendipitously of the same "mind" in what concerns the poetics informing our approach, ChatGPT also seemed to tackle the assignment—in both English and Romanian—with a context-and-corpus-aware method. We did ask the LLM whether it was indeed aware of the *Yale Anthology* and its contents, and its response was quite politic. It knew *of* it but in its capacity as "AI language model" it did not have access "to information about the specific contents of my training data," while being nevertheless "familiar with a wide range of poetry and literary works, including those by many French poets featured in such anthologies…" (ChatGPT 2024)

Our next experiment was then to create a graph of the poems in all three languages, hence a 6-layered network with 2 layers (tf-idf and enjambment similarities) per language, conventionally considering poems and their translations to be the same topological entities, and thus obtaining a 588-node multiplex. The behaviour of the 3 monolingual sets (and multiplexes) separately and that of the trilingual set (and overall network) are in this particular case similar in certain respects and different in significant others. There are 23 titles that made the cut in the trilingual multiplex which is also the sum of all titles in the three languages developed separately, which may lead to the wrong conclusion that in terms of

numbers the two models behave very similarly. This represents nevertheless a manner of speaking when it comes to the trilingual model, conventionally considering a poem to be the same as its translations, when, in reality, we have 69 poems in French, English, and Romanian ranking among the first 3 x 69 (207) of the 3 x 98 (294) total in all three languages combined. For the single-language models, the 23 total cannot be multiplied by 3 though. Of these 23, as stated above, only 18 make the cut across the languages modeled separately, therefore, only 54 (compared to the 69 in the trilingual model) will rank on top of the 294 in all three languages. In spite of the relatively small sizes of the corpora under discussion, the advantage of complexity and multilingualism is obvious in this case when we compare the two models, a point we made in other recent related publications with regards to community-based resistance (in interplatform intermediality and data-commoning, Tanasescu 2024a) and empowering minorities and minoritized poetries (Tanasescu 2024b).

In terms of specific differences, the one-language-at-a-time model managed to insert among the top-ranking titles five that the trilingual did not: "Vive la Bauté" by Laurence Vielle, "Brux'elle" by Anne Versaille, "Desincarceration" by Charline Lambert, and "Le P Respectueux" and "Le Coiffeur," both by Jan Baetens, listed here with their original titles while they actuyally made the cut in each of the three languages (see again the paper's GitHub). These poems are united by an ambiguity and/or political ambivalence in representing femininity that ranges from the discriminating and commodifyingly normative in the first one to the social/ sexual/work-related in the second—as well as in Baetens's both poems—to the medical-physiological in Lambert's "Desincarceration." On the other hand, the trilingual model's specific contribution consists of "Ici" (part II) by Corinne Hoex, "1998" by Lisette Lombé, "Mains Ouvertes" (parts I and II) by Philippe Mathy, and "Promener Un Feu" by Carl Norac. A rich and nuanced stylistic and thematic spectrum therefore ranging from Hoex's delicate, catalog-like, yet intransigently materialist take on the established histories of the spiritual, the incisive performative assertion of the underrepresented or discriminated in Lombé to the alluringly self-contradictory, both pregnant and airy, erotic and nature-contemplating lyricism of Mathy to Norac's ironically sentimental and matter-of-fact philosophical allegories.

These two trends are fused in the 18 titles (translating to 54 poems in the three languages) that are shared by both models as ranking high in the enlarged corpora, with a further stress on the feminine, the underrepresented, and the political, as complemented by the erotically intense or contemplative. Both poems by Aventin are in there, for instance, alongside those of Véronique Bergen and Gioia Kayaga, and so are Hoex and Lambert's poems in the Belgian corpus that did not appear in each model's specific lists above (see MARGENTO 2024B). Savitzkaya's profane and infectious eroticism alongside Namur's quintessential music and Logist's heartfelt, inconsolate paradoxes complete that large, diverse, and commanding ensemble.

Poetics of the Technics. In Guise of a Conclusion

We think of anthology editing as a computational way of revisiting and redeveloping poetry corpora and their contents. For our first experiment, for instance, we constructed three multiplexes, one for each language, the first comprising the originals, the second the English translations by "humans" in the *Yale Anthology* alongside ChatGPT's translations into English of the Belgian corpus, and the third the DeepL translations into Romanian of the French poems in the anthology alongside ChatGPT's Romanian translations of the Belgian corpus. In this case, our resulting "imaginary" anthology would first include the Belgian poems that made the cut in all three languages and would therefore be featured as such, in French alongside English and Romanian translations. Then it would include the ones that made it only in French—and those one would be featured in French only—and, last but not least, the one poem that made it in Romanian alone, featured as such, only in Romanian.

This involves not only reshuffling the contents of the *Yale Anthology* and replacing a significant number of poems with Belgian ones, but also two more important steps: first, assessing the stylistic prominence of poems in different languages and, where a certain language made a significant contribution, featuring the poem as such (specifically in that language) in the selection (featuring, as it were, the… original translation); and, second, rewriting the initial corpus itself.

On the one hand, given the computational processing and the Mathematical model, the poems are reread and rewritten as networked agents: their presence in the multiplex creates new (inter/intra)relational versions of all the other poems (by changing their tf-idf vectors for instance, or their topological location in the resulting complex network) which, in their turn, read and rewrite the poems that triggered such an evolution. Translations of the newly added Belgian poems thus make their way into an expanding more inclusive corpus while new translations of some of the poems in the initial one replace those initially featured (in, e.g., the *Yale Anthology*) and/or complement them with other translations, into the same and other languages. On the other hand, the larger project this came as a development of, involved translation as rewriting and/or rearranging poems in the initial and overall corpora. For the Belgian anthology (an anthology of possible anthologies, as described in a theoretical section of the forthcoming collection, Tanasescu 2024b), some of the poems in the *Yale Anthology* (particularly in the most recent-period section) since prose poems were rearranged into verse with a sharp eye on possible enjambments to crop up in the process as arguably both consistent with and enhancing of the piece's as well as the corpus' semantic and aesthetic potential. These experiments were once in a while conducted in collaboration with ChatGPT and other LLMs and while mainly dealing with prose poems, it also now and then involved poems in verse in both the Yale and the Belgian corpora (MARGENTO 2024). Such experiments are currently under way in the trilingual corpora too and will constitute the focus of future publications.

There are two concepts we want to advance here based on the technological poetics informing the project (and related projects) and the attendant computational experiments and outputs. The first of them is *iterative corpus translation.* The poetry corpora involved and created (in and by translation) translate, reshuffle, inter-penetrate each other and intermingle by means of NLP and network assemblage and analysis. Poems are translated within, across, and as corpora into the same and other (natural and computational) languages and forms iteratively and the output is itself relational, emergent, dynamic, inter/multilingual, in (potentially) continuous motion and change, and further translatable (or dissolvable) into ever wider contexts and environments. This modeling of translation aligns with the non-linear, rhizomatic, performative, and ludic dimension that has become so attractive in the contemporary

study of translation. And it does so without compromising on the communicative and descriptive model of translation, or on its radical, political, and social implications: the iterative corpus translation places underrepresented poets and poems in a dialog with so-called mainstream one and emphasizes their transformative role in larger and larger amounts of data.

Second, there is the concept of *translation modeling.* Translation involves, when it is not in and of itself, a form or an array of forms of modeling. In poetic-algorithmic and computational translation though, the translation of models and, moreover, the translation of modeling functions, is also at play. As we have seen in our experiments for instance, monolingual corpus multiplexes were translated into sets of inter-related multiplexes, and later on, into an overall trilingual multiplex. The model not only gets translated into other (evolving) models, but also undergoes translation as its implementation reshapes the data while the latter in its turn (re)shapes the model in evolving nested environments. Quantifying poetic features in several languages simultaneously prompted, in our case, conventionally deeming the nodes to be the same in each language multiplex, which in its turn made possible a larger more inclusive multilingual model. Paradoxically though, this actually triggered the emergence of each language (in translation) as distinctive and irreplaceable constituent of the model and modeling actor in its own right. It also opened the door to diversity at scale, as basically any element of any corpus, and thus any corpus really, became replaceable or hybridizable by multiple translations. The line between "original" and translation not only gets blurred, but can disappear altogether as source and target corpora (and languages) shape each other interactively and, in fact, can only be defined, (co-)created, and read as interrelatedly performative. Translation modeling consequently not only involves the modeling of the process of translation but also modeling in or by means of translation.

Therefore, while our concept can be obviously read as modeling translation in/as translating models (or, rather, translating modeling) it also comports the experimental elucidation of a theoretical notion. Modeling—as fundamentally performative, intermedial, and political (see the concept of "inter*model*ity," Tanasescu 2024a)—and translation—which also intimately shares the same features—inform each other on various levels, some of the most intriguing of which

can become apparent in computationally assembled poetry anthologies. The act of modeling and translating becomes not merely a technical process but a creative and transformative endeavor. Through computational methods, poetry anthologies can be curated and constructed to showcase how different linguistic and cultural elements interact and influence one another, shedding light on the performative and intermedial aspects inherent in both modeling and translation as well as on the hypercyborg, collage, avantgarde, playful, performative, and transmedial translator.

Annex 1: Figures

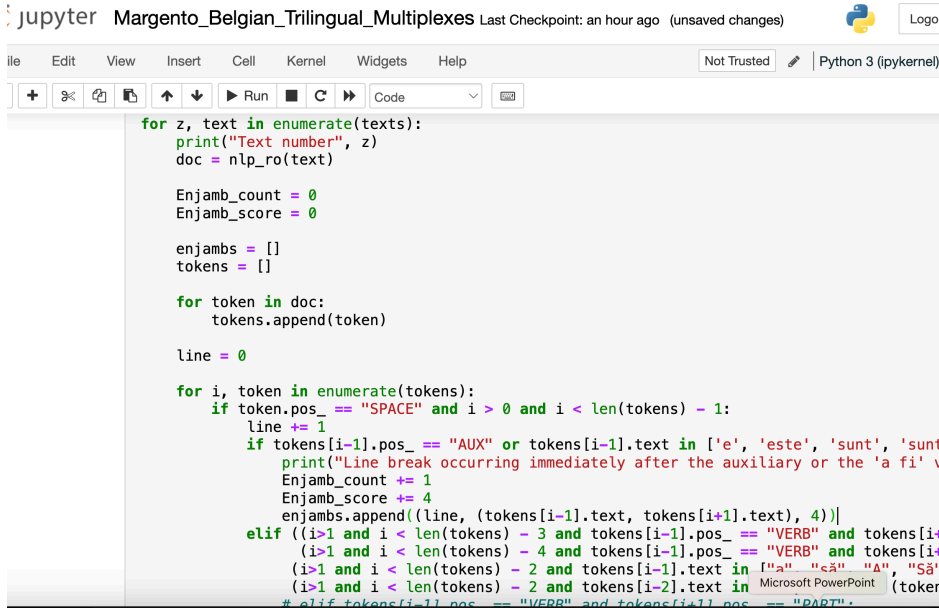


Figure 1: A screenshot of the Python script quantifying enjambments in Romanian

References

Armaselu, Florentina. 2024. "Hermeneutic Modeling of Detail in Textual Zoom and Literary Texts." In *Literature and Computation. Platform Intermediality, Hermeneutic Modeling, and Analytical-Creative Approaches*, edited by Chris Tanasescu, 119–146. New York; London: Routledge.

Bennett, Karen. 2024. "Epistemic Translation: Towards an Ecology of Knowledges." *Perspectives Studies in Translation Theory and Practice*. https://doi.org/10.1080/0907676X.2023.2294123.

Boria, Monica, Ángeles Carreres, María Noriega-Sánchez, and Marcus Tomalin (eds). 2020. *Translation and Multimodality. Beyond Words*. New York: Routledge.

Caws, Mary Ann (*ed*). 2008. *The Yale Anthology of Twentieth-Century French Poetry*. New Haven, CT: Yale University Press.

Chesterman, Andrew. 2000. "A Causal Model in Translation Studies." In *Intercultural Faultlines. Research Models in Translation Studies: v. 1: Textual and Cognitive Aspects*, edited by M. Olohan, 15–27. New York: Routledge.

Fabo, Pablo Ruiz, Clara Martínez Cantón, Thierry Poibeau, and Elena González-Blanco. 2017. "Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets." In *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 27–32. Vancouver, Canada: Association for Computational Linguistics.

Frawley, William. 1984. "Prolegomenon to Translation." In *Translation: Literary, Linguistic, and Philosophical Perspectives*, edited by William Frawley, 159–178. Newark: University of Delaware Press.

Goldstein, Jeffrey. 1999. "Emergence as a Construct: History and Issues." *Emergence* 1 (1): 49-72. https://doi.org/10.1207/s15327000em0101_4.

Grass, Delphine. 2023. *Translation as Critical-Creative Practice*. Cambridge: Cambridge University Press.

Guerberof-Arenas, Ana, and Antonio Toral. 2022. "Creativity in Translation. Machine Translation as a Constraint for Literary Texts." *Translation Spaces* 11 (2): 184–212.

Hadley, James, Kristiina Taivalkoski-Shilov, Carlos Teixeira, and Antonio Toral. 2022. *Translation Technologies for Creative-Text Translation*. London; New York: Routledge.

Kesarwani, Vaibhav, Diana Inkpen, and Chris Tanasescu. 2021. "#GraphPoem: Automatic Classification of Rhyme and Diction in Poetry." *Interférences littéraires / Literaire interferenties* 25, edited by Chris Tanasescu, 218–235. https://bit.ly/ 3JUhEla.

Kesarwani, Vaibhav, Diana Inkpen, Stan Szpakowicz, and Chris Tanasescu. 2017. "Metaphor Detection in a Poetry Corpus." In *Proceedings of the Annual Meeting of the*

Association for Computational Linguistics 2017. Vancouver, Canada. http://bit.ly/2BX0lOO.

Larrue, Jean-Marc, and Marcello Vitali-Rosati. 2018. *Media Do Not Exist. Performativity and Mediating Conjunctures*. Translated by John Detre and Beth Kearney. Amsterdam: The Institute of Network Cultures.

Latour, Bruno. 1993. *We Have Never Been Modern*. Translated by Catherine Porter. Cambridge: Harvard University Press.

Lee, Tong-King. 2002. *Translation as Experimentalism: Exploring Play in Poetics*. Cambridge: Cambridge University Press.

Leithauser, Brad. 2022. *Rhyme's Room. The Architecture of Poetry*. New York: Knopf.

Longenbach, James. 2007. *The Art of the Poetic Line*. Minneapolis: Graywolf Press.

Loffredo, Eugenia, and Manuela Perteghella (eds). 2006. *Translation and Creativity. Perspectives on Creative Writing and Translation Studies*. New York: Continuum.

Lou, Andres, Diana Inkpen, and Chris Tanasescu. 2015. "Multilabel Subject-Based Classification of Poetry." *FLAIRS 28 (Proceedings of the 28th Florida Artificial Intelligence Research Society Conference)*, 187-192. http://bit.ly/2CmPvCK.

Lukes, Alexandra (*ed*). 2023. *Avant-Garde Translation*. Leiden: Brill.

Malmkjær, Kirsten. 2019. *Translation and Creativity*. New York: Routledge.

MARGENTO (*ed*). 2018. *"US" Poets Foreign Poets / Noi poeți americani poeți străini*. Bucharest: Fractalia.

MARGENTO (*ed*). 2024a. *A Computationally Assembled Anthology of Contemporary Belgian Poetry*. New York: Peter Lang. Forthcoming.

MARGENTO. 2024b. *PoetryTranslationModeling* (GitHub repo). https://github.com/Margento/PoetryTranslationModeling.

Robert-Foley, Lily. 2024. *Experimental Translation. The Work of Translation in the Era of Algorithmic Production*. London: Goldsmith Press.

Robinson, Douglas. 2022. *The Experimental Translator*. Cham: Springer.

St. André, James. 2023. "Implications of Computer Code Translation for Translation Studies." *Translation Studies* 16 (2): 195–211. https://doi.org/10.1080/14781-700.2023.2205418.

Stange, Jan-Erik. 2024. "Reflective Modeling (Modeling What Would Be There): A Critical, Creative, and Constructive Approach to Data Modeling." In *Literature and Computation. Platform Intermediality, Hermeneutic Modeling, and Analytical-Creative Approaches*, edited by Chris Tanasescu, 173–210. New York; London: Routledge.

Steiner, George. 1992. *After Babel: Aspects of Language and Translation*. 2nd ed. Oxford: Oxford University Press.

Tanasescu, Chris. 2024a. "Dynamical Systems and Interplatform Intermediality. The Case of #GraphPoem @ DHSI." In *Literature and Computation. Platform Intermediality, Hermeneutic Modeling, and Analytical-Creative Approaches*, edited by C. Tanasescu, 25-58. New York; London: Routledge.

Tanasescu, Chris. 2024b. "Introduction." In *Computationally Assembled Anthology of Contemporary Belgian Poetry*, edited by MARGENTO. New York: Peter Lang. Forthcoming.

Tanasescu, Chris (*ed*). 2024c. *Literature and Computation. Platform Intermediality, Hermeneutic Modeling, and Analytical-Creative Approaches*. New York; London: Routledge.

Tanasescu, Chris. 2022. "#GraphPoem @ DHSI. A Poetics of Network Walks, Stigmergy, and Accident in Performance." *IDEAH Journal* 3 (1). https://ideah.pubpub.org/pub/tgdpsell/release/4.

Tanasescu, Chris. 2011. "The Graph Poem (Four Poems & a Note on the Poetics)." *Poems and Poetics*. https://poemsandpoetics.blogspot.com/2011/03/chris-tanasescu-graph-poem-four-poems.html.

Tanasescu, Chris, and Raluca Tanasescu. 2023. "Complexity and Analytical-creative Approaches at Scale. Iconicity, Monstrosity, and #GraphPoem." In *Zoomland. Exploring Scale in Digital History and Humanities*, edited by Florentina Armaselu and Andreas Fickers, 237-260. Berlin: De Gruyter. https://doi.org/10.1515/ 9783111317779-010.

Tanasescu, Chris, Diana Inkpen, Vaibhav Kesarwani, and Prasadith Kirinde Gamaarachchige. 2020. "A-Poetic Technology. #GraphPoem and the Social Function of Computational Performance." *DH Benelux Journal* 2: 1–14. https://journal.dhbenelux.org/journal/issues/002/article-39-tanasescu/article-39-tanasescu.html.

Tanasescu, Chris (MARGENTO), Vaibhav Kesarwani, and Diana Inkpen. 2018. "Metaphor Detection by Deep Learning and the Place of Poetic Metaphor in Digital Humanities." *FLAIRS 31: Proceedings of the 31st Florida Artificial Intelligence Research Society Conference*, 122–127. https://www.aaai.org/Library/FLAIRS/ flairs18contents.php.

Tanasescu, Chris, Bryan Paget, și Diana Inkpen. 2016. "Automatic Classification of Poetry by Meter and Rhyme." In *FLAIRS 29: Proceedings of the 29th Florida Artificial Intelligence Research Society Conference*, 244–249. http://bit.ly/2FAzdoA.

Toury, Gideon. 1995. *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Vermeer, Hans. 1996. *A Skopos Theory of Translation: Some Arguments for and against*. Heidelberg: TEXTconTEXT Verlag.

Weblinks

ChatGPT (2024) "Feu de l'anniversaire." Chat with Chris Tanasescu. March-April 2024.

Gemini (2024) April 2 chat with Chris Tanasescu.

Challenges of Digitization and Digitalization for the Study of Writing: from archival survey to process analysis

Georgeta CÎȘLARU

This article looks at the impact of digitization and digitalization on the study of writing from an interdisciplinary perspective, with a specific focus on *keylog* data. According to Brennen and Kreiss (2016), *digitization* can be defined as the material process of converting analog streams of information into digital bits, and *digitalization* as the way domains of social life are restructured around digital communication and media infrastructures. The former is necessarily determined by the latter, with tools and practices mutually conditioning each other.

The impact of digitization on the very modalities of research in the human sciences is widely recognized (Dalbello 2011). Digital Humanities are changing the epistemologies and ontologies underlying research (Berry 2012). One of the most obvious impacts on humanities research, of which a large part involves analyzing and interpreting written texts (McGillivray *et al.* 2020), is the volume of archives that can be processed thanks to digitization on the one hand, and the automation it enables on the other (Moretti 2007).

Since its emergence as a technology (Auroux 1994), writing has undergone many changes, evolving from physical to digital media. Recent publications have examined this evolution and highlighted the many changes it has entailed (Van Waes *et al.* 2006; Kruse *et al.* 2023). These changes range from the use of word processing software (Heilmann 2023) to the development of digital tools for writing (argumentation, note-taking, etc., e.g., Benetos 2023), learning and regulation (plagiarism, self-reflexivity, e.g., Devitt *et al.* 2023). Digital technologies have not only impacted the ways of writing and favored new discourse genres (Rowley-Jolivet and Carter-Thomas *in print*), they have also helped to develop new methodologies for investigating writing and the writing process (Van Waes *et al.* 2011).

Impacts on the nature of the data

Archives in writing studies

For some disciplines in the humanities, access to archives is a prerequisite for research. In the field of writing studies, the literature has focused on the evolution of writing systems, from a phylogenetic perspective, or on texts in the making, from an ontogenetic perspective. The latter is at the heart of critical genetics (Hay 2002).

The anteriority of the written text has also interested linguists in its variational, dynamic, and enunciative compositional dimension, shifting the research focus from the text entity to the act or process of production, writing, observed from drafts. Since the 1970^{s}, critical genetics has provided a vast body of work on the archaeology of texts, studying the traces of the writing process (*cf.* Hay 2002, Grésillon 2016 [1994]). Genetic criticism, which was first interested in literary creativity, happy accidents and the protean milestones of writing (*cf.* Grésillon 2008, 11), places at the center of its analysis erasure, reformulation and other operations that ultimately constitute the four operations of rewriting: addition, deletion, replacement and displacement (Grésillon 1989). These operations are all traces that reconstitute entire sections of the process: for instance, "erasure is one of the elements capable of bearing witness to the temporal dimension specific to any writing process" (Grésillon 2016 [1994], 83). The identification by textual genetics of these four operations of text revision introduces a distributional perspective that opens the way to multiple analyses. Linguistic issues were rapidly identified and have been the subject of several development proposals (*cf*. Grésillon and Lebrave 2008).

Attempts have been made to describe types of writers according to the way they use (or do not use) drafts. Some authors consider the draft itself and the "drafting" activity as part of the process of idea generation and organization, that is, planning (Sharples and Pemberton 1990). The aim is to understand the impact of different types of drafts (lists of words, plans, smoothed drafts, etc.) on the writing itself and on the quality of the text produced (Piolat and Roussey 1996). Other authors identify types of writers according to the more or less planned or programmatic nature of their drafts. Galbraith (1992), for example, distinguishes between "executor" writers (who base their

drafting on highly organized drafts) and "discoverer" writers (who write from very sketchy, poorly organized drafts, generating ideas and their plan essentially as they write). Fenoglio (2007) points out that this distinction between programmatic and free writing seems rather random, apart from the typical cases of Zola and Proust, respectively executor and discoverer.

Whether the focus is on revision operations, or on the complex dynamics of cognitive processing underlying written production, everything concurs in considering the textualization process in a non-linear way. In genetic criticism studies, the observation of rewriting operations, various graphic marks on drafts and the analysis of conceptual elaborations also show the non-linearity of the writing process (for recent work, *see* Fenoglio 2015). The text is conceived in terms of deployment, building on pre-existing elements (the pre-text, made up of memorized information, standards, formats) and with elementary units such as keywords, proto-words, etc., which are then used to create the text, as shown by *Figure 1*.

Born-digital archives and the writing process

While there is no consistent evidence of the benefits of replacing handwriting by typing in early writing instruction (Wollscheid *et al.* 2016), the use of digital tools to compose a text has impacted writing practice. Keyboard writing in itself is the exact opposite of handwriting, as it operates with discrete script signs, which may influence the apprehension of linguistic units: letters, words, and phrases (Doquet 2006). Physical space management is also different, with text displayed in small windows on the screen, and revision traces are evanescent (Anis 1995).

The effects on the process and the product of writing are diverse, but in all cases, social practice in itself and the object of study are affected, as reported by various studies. Whilst reading on paper shows more positive affordances for Finnish college students, digital writing is valued because it enables the text to be quickly and efficiently edited (Taipale 2014). Williams and Beam (2019) point out in their review that digital technologies and writing tools increase motivation and social collaboration in students. Dahlström and Boström (2017) showed that university students produce longer and linguistically more correct narrative texts on tablets compared to handwritten texts. They also noticed differences in representing

processes, with material processes dominating in digital writing, and material and relational processes dominating in handwriting.

Compared to handwritten or typewritten data, texts produced by means of digital tools are formally and qualitatively different, and offer different research perspectives. Drafts become evanescent and not archival, and revisions leave no visible physical trace. The use of digital tools also offers new resources for scientific investigation, however, and new solutions for accessing the virtual "traces" of writing, for instance by recording the progressive construction of the text through its different versions. It is then possible to run linguistic processing (automatic or otherwise) on the raw texts collected directly. First and foremost, this allows longitudinal approaches in short diachrony, by comparing the versions of texts from each writing session and identifying variants that interfere with the linearity of the texts. We call this macro-segmentation. In this way, textual genetics can calculate additions, replacements and deletions to obtain an overall view of the modifications a text has undergone. Software such as EDITE MEDITE6 (Fenoglio and Ganascia 2007) or ALLONGOS (Lardilleux *et al.* 2013) compare each new text version with the previous one. For example, ALLONGOS compares versions and indicates which revision operations have been carried out.

The various writing sessions reveal stabilized versions of the text in the making and constitute traces of the stages of discursive configuration. Ferrer (2011, 182) insists on the role and cognitive interest of drafts for textual configuration, considering variants as "protocols for making of a text".

Process analysis

The "protocols for making a text" have always intrigued researchers seeking to understand the cognitive mechanisms of writing. Previously, researchers exploited verbal protocols in order to make mental mechanisms "observable" by asking individuals to describe aloud the activities they performed, either concomitantly or retrospectively (Levy and Ransdell 1995, 1996). These protocols led to models of written communication distinguishing planning (cognitive activities: activation-selection, organization), writing (transcription, translation: verbalization, from thought

to language), and revision (reading-evaluation-correction) (Flower and Hayes 1980; Hayes and Flower 1980).

These reflective approaches were intended to compensate for the lack of access to the writing process. In the 1990s, the evolution of digital tools made it possible to access process information through screen capture, then by means of tablets, screen-capture or key-stroke logging tools. These developments led to the study of eye movements and grapho-motor activity during writing on Eye-and-Pen tablets (Chesnet and Alamargot 2005), the study of L2 writing in university students via the screen capture tool Screencast-O-Matic (Seror 2013), process recording and analysis via keylogs such as Inputlog (Leijten and van Waes 2006) or Scriptlog (Strömqvist and Malmsten 1998). Wengelin and Johansson (2023) offer a detailed description of the tools and their functionalities.

A chrono-topological longitudinal approach to writing and textualization

Key-stroke logging software provides a chronologically structured record of all the operations involved in writing on a keyboard: typing characters and spaces, revisions (deletions, insertions, moves, etc.), pauses sequencing the process, or consultation of sources outside the document. All these data can be updated in their longitudinal dynamics thanks to their rereading function, extracted by type (pause, typing, revision), classified by series, etc. Recorded writing has thus become a means and an object of analysis, in cognitive psychology, through experimental studies, but also in communicative approaches, such as the pragmatics of writing interested in progress analysis (*see*, for example, Perrin 2013). By recording and analyzing chronometric data associated with production, such as pauses or writing speed, it is possible to attempt to infer the cognitive mechanisms at play. *Pausology*—the study of pauses in language production—is an essential part of any approach to the processual dimension of language production. Pauses provide a first level of segmentation of language flow, whether oral or written, and enable us to associate the production of textual units with the dynamics of the writer's or speaker's mental functioning (*cf.* Fayol 1997).

By offering longitudinal access to the chronological unfolding of the writing process, keystroke recording tools stage a representational paradox. While the studies

based on material archives and drafts demonstrate the non-linearity of the writing process, the temporality of recorded writing is always linear. Only by relating this chronological linearity to the topology of the text is it possible to articulate temporality to the non-linearity of the process. For example, graphic representations of the writing process proposed by Leblay and Caporossi (2015) clearly illustrate the recursivity of revision operations, and Inputlog graphically represents the progress of writing in terms of characters typed or deleted, pauses, and sources consulted (*Figure 3*).

In addition to the writer's actions on the text, *Inputlog* software records all the writer's actions on the computer, outside the word processor. This means that, while writing, it is possible to consult various files stored on the computer, go on the Internet, copy text, write and send an e-mail, etc. All this information is recorded, making it possible to analyze the sources consulted by the writer. With all this information recorded, it becomes possible to analyze the sources consulted by the writer and see whether they have been integrated into the text. The software can also encode metadata to link a writing process to a specific writer, or compare data from different writers. It can also use the generic file created to perform a number of analyses of the writing process. *Inputlog* provides information on texts written (*i.e*., number of words, sentences, words per sentence in the final text or during the session), on revision operations (by creating files summarizing the various operations carried out), on pauses (i.e., number, average duration, pauses between words, sentences, paragraphs) and execution periods (see below). It also calculates keystroke metrics to analyze the typing skills of writers.

The identification of textual operations performed by writers when composing their text is based on S-notation, a language developed specifically to support the analysis of revision episodes (Severinson Eklundh and Kollberg 1996). The software also enables a wide range of analyses of the writing process to be carried out automatically. From this point of view, digital development affects not only the nature of the archive (born-digital, the digital edge) and the point of view on data (processual), but also the place of the analyst, by offering him or her immediate avenues of observation.

Linguistic insights

In the field of linguistics, two methodological questions emerge as the object itself mutates: the first concerns the flow of language data recorded in real time, and raises the question of what can be considered the object of linguistic analysis, *i.e.*, how to discriminate between "analyzable" linguistic units? The second concerns the transition from the language sequence level to the text level, *i.e*., how analysts recover the text as a product when moving away from it both upstream in time and with regard to its unitary structure.

Pauses in the writing process can be considered as segmentation markers, according to which it is possible to extract and categorize segments of linguistic production. Pauses segment textual sequences in the process of elaboration. These sequences can be considered to have the status of linguistically attested empirical data. We can define pausing as spontaneous segmentation, or auto-segmentation, insofar as pauses are the work of the writers and not of the analyst or the tool, which merely records and measures them. From this perspective, the primary object of analysis we focus on is linguistic sequences written uninterruptedly between two pauses in production. We refer to these sequences as *bursts of writing*, a concept borrowed from the psycholinguistics of writing (*cf*. Chenoweth and Hayes 2001) and reworked in the frame of linguistics by Cislaru and Olive (2018). From a linguistic point of view, we can thus consider textual bursts as units of written production, or units of performance.

The duration of pauses, the length of the bursts, and the way in which pauses and bursts are structurally arranged constitute ecological data that can be studied in their own right. This new form of segmentation does not compete with all the other forms of segmentation used to analyze texts and discourse (repeated segments, sequences, periods, motifs, formulas, etc.), nor does it represent an exclusive alternative, but rather a temporal and processual approach to text and writing (Cislaru and Olive 2018).

This processual view of writing brings it closer to oral production data, implying similar questions about flow segmentation and the status of the pause (Grosman *et al.* 2018). In this respect, we can consider that we are dealing with a new epistemological mutation engendered by the evolution of digitalization.

Last but not least, on-line revisions reveal metadiscursive approaches whose traces are undetectable in the final product and which are sometimes identifiable only

through the movement that consists in filling a digital space with two different forms successively. Yet the relationship (which it might be questionable to call "equivalence", as in reformulations) between these forms and between the chosen form, the text in itself and the writing process remains to be clarified: should these forms be considered revisions, or on the contrary production, if they impact the evolution of the text? These issues led to new classification suggestions (Conijn *et al.* 2022) and show that revisions may be considered as process-oriented or product-oriented.

Conclusion

This article takes advantage of a brief historical overview of studies on the writing of texts to show the contributions, implications and mutations of digital tools in the study of writing, moving from the physical archive to the born-digital archive, from the sequential study of the process to on-line study. This evolution shows the transformation of the object of study itself, with the new issues it implies and the fields of research that are being redrawn.

Annex 1: Figures

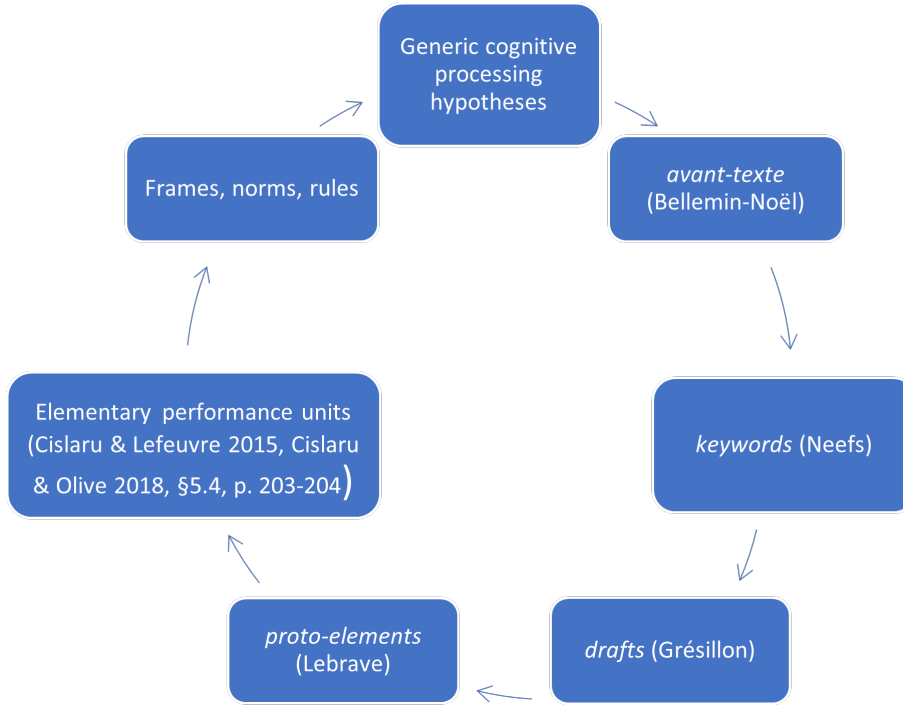


Figure 1: Cognitive-linguistic text deployment.

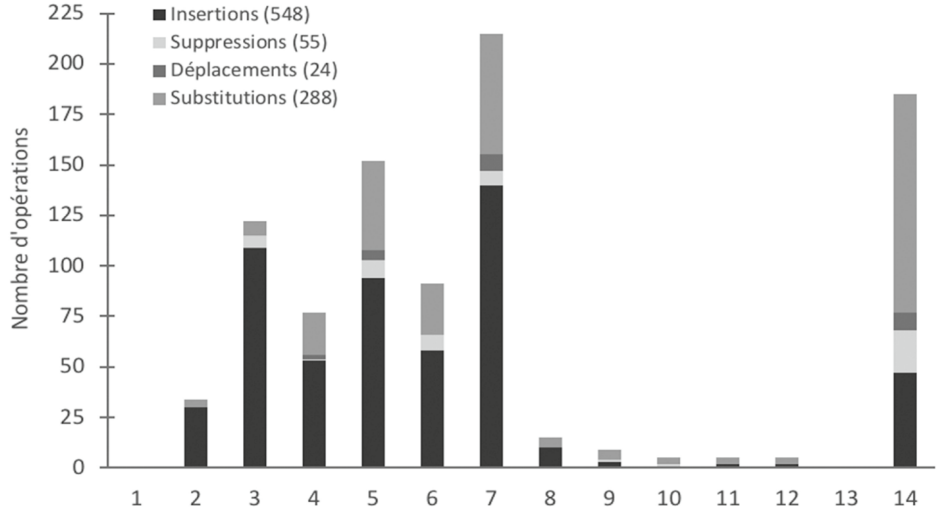


Figure 2: Number of rewriting operations for a 14-session writing process.

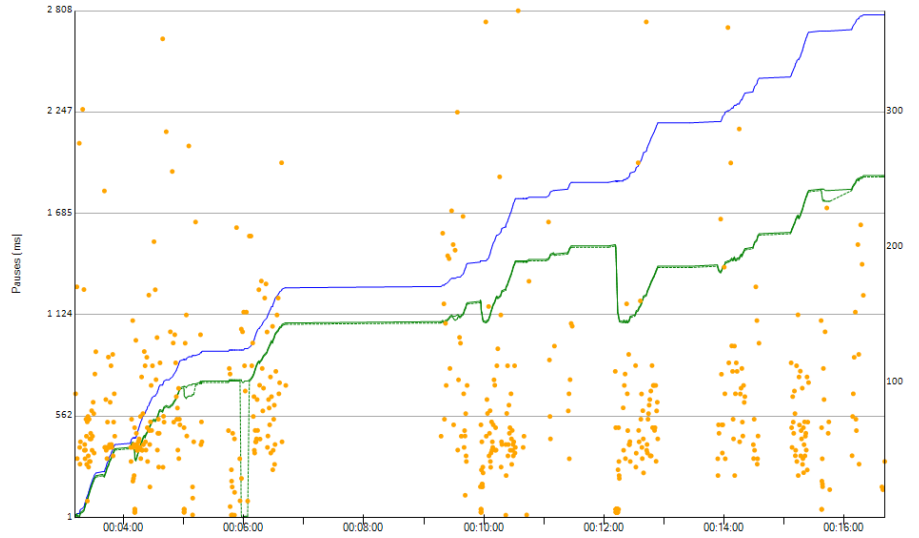


Figure 3: Inputlog representation of a writer's productivity during a social report writing session.

The yellow dots indicate pauses, each of whose duration (in ms) is shown on the left vertical axis, with the moment of occurrence indicated on the horizontal axis. The green curve represents the sum of characters present in the text. The blue curve indicates the cumulative sum of characters typed since the start of the writing session. A descending green curve indicates the deletion of characters. The dotted green curve indicates the cursor position (Cislaru and Olive 2018).

References

Anis, Jacques. 1995. “Le traitement de texte: écriture ou méta-écriture?” *Repères* 11: 15–27.

Auroux, Sylvain. 1994. *La révolution technologique de la grammatisation*. Liège: Mardaga.

Bellemin-Noël, Jean. 1972. *Le texte et l’avant-texte*. Paris: Larousse.

Benetos, Kalliopi. 2023. “Digital Tools for Written Argumentation.” In *Digital Writing Technologies in Higher Education. Theory, Research, and Practice*, edited by Otto Kruse, Christian Rapp, Chris M. Anson, Kalliopi Benetos, Elena Cotos, Ann Devitt, and Antonette Shibani, 81–99. Springer.

Berry, D. M. 2012. “Introduction: Understanding the Digital Humanities.” In *Understanding Digital Humanities*, edited by D. M. Berry, London: Palgrave Macmillan. https://doi.org/10.1057/ 9780230371934_1.

Brennen, J. Scott, and Daniel Kreiss. 2016. “Digitalization.” In *The International Encyclopedia of Communication Theory and Philosophy*, 1–11. John Wiley & Sons.

Chenoweth, Norton Ann, and John R. Hayes. 2001. “Fluency in Writing: Generating Text in L1 and L2.” *Written Communication* 18 (1): 80–98.

Chesnet, David, and Denis Alamargot. 2005. “Analyse en temps réel des activités oculaires et grapho-motrices du scripteur: intérêt du dispositif ‘Eye and Pen’.” *L’année psychologique* 105 (3): 477–520.

Cislaru, Georgeta, and Florence Lefeuvre. 2015. “The Instrumental Use of Verbless Sentences in Writing and Rewriting: A Longitudinal and Genre-Contrasted Point of View.” In *Writing(s) at the Crossroads: The Process/Product Interface*, edited by Georgeta Cislaru, 57–80. Amsterdam–Philadelphia: John Benjamins.

Cislaru, Georgeta, and Thierry Olive. 2018. *Le processus de textualisation*. Bruxelles: De Boeck.

Conijn, Rianne, Emily D. Speltz, Menno van Zaanen, Luuk van Waes, and Evgeny Chukharev-Hudilainen. 2022. “A Product- and Process-Oriented Tagset for Revisions in Writing.” *Written Communication* 39 (1): 97-128. https://doi.org/10.1177/074108832-11052104.

Dahlström, Dahlström, and Boström, Boström. 2017. “Pros and Cons: Handwriting Versus Digital Writing.” *Nordic Journal of Digital Literacy* 12 (4): 143–161.

Dalbello, Marija. 2011. “A Genealogy of Digital Humanities.” *Journal of Documentation* 67 (3): 480–506.

Devitt, Ann, Kalliopi Benetos, and Otto Kruse. 2023. “Writing and Learning: What Changed with Digitalization?” In *Digital Writing Technologies in Higher Education. Theory, Research, and Practice*, edited by Otto Kruse, Christian Rapp, Chris M. Anson, Kalliopi Benetos, Elena Cotos, Ann Devitt, and Antonette Shibani, 423–433. Springer.

Doquet-Lacoste, Claire. 2006. “L’objet insaisissable: l’écriture sur traitement de texte.” *Genesis* 27: 35–44.

Fayol, Michel. 1997. *Des idées au texte. Psychologie cognitive de la production verbale, orale et écrite*. Paris: Presses Universitaires de France.

Fenoglio, Irène. 2007. "Du texte avant le texte. Formes génétiques et marques énonciatives de prévisions textualisantes." *Langue française* 155: 8–34.

Fenoglio, Irène. 2015. "From Writing Under Production to the Finished Product: A Processual Threshold." In *Writing at the Crossroads: The Process-Product Interface*, edited by Georgeta Cislaru, 25–50. Amsterdam – Philadelphia: John Benjamins.

Fenoglio, Irène, and Jean-Gabriel Ganascia. 2007. "MEDITE: un logiciel pour l'approche comparative de documents de genèse." *Genesis* 27: 166–168.

Ferrer, Daniel. 2011. *Logiques du brouillon. Modèles pour une critique génétique*. Paris: Éditions du Seuil.

Flower, Linda S., and John R. Hayes. 1980. "The Dynamics of Composing: Making Plans and Juggling Constraints." In *Cognitive Processes in Writing*, edited by Lee W. Gregg and Erwin R. Steinberg, 31–50. Hillsdale, NJ: Lawrence Erlbaum.

Galbraith, David. 1992. "Conditions for Discovery Through Writing." In *Computers and Writing. Issues and Implementations*, edited by Mike Sharples, 45–72. Dordrecht: Kluwer Academic Publishers.

Grésillon, Almuth. 1989. "Fonctions du langage et genèse du texte." In *La naissance du texte*, edited by Louis Hay, 177–192. Paris: José Corti.

Grésillon, Almuth. 2008. *La mise en œuvre. Itinéraires génétiques*. Paris: CNRS Éditions.

Grésillon, Almuth. 2016 [1994]. *Eléments de critique génétique*. Paris: CNRS Éditions.

Grésillon, Almuth, and Jean-Louis Lebrave. 2008. "Linguistique et génétique des textes: un décalogue." *Le français moderne*, numéro spécial "Tendances actuelles de la linguistique française": 37–49.

Grosman, I., A. C. Simon, and L. Degand. 2018. "Variation de la durée des pauses silencieuses: impact de la syntaxe, du style de parole et des disfluences." *Langages* 211: 13–40.

Hay, Louis. 2002. *La littérature des écrivains. Questions de critique génétique*. Paris: Corti.

Hayes, John R., and Linda S. Flower. 1980. "Identifying the Organization of Writing Processes." In *Cognitive Processes in Writing*, edited by L. W. Gregg and E. R. Steinberg, 3–30. Hillsdale, NJ: Lawrence Erlbaum.

Heilmann, Till A. 2023. "The Beginnings of Word Processing: A Historical Account." In *Digital Writing Technologies in Higher Education. Theory, Research, and Practice*, edited by Otto Kruse, Christian Rapp, Chris M. Anson, Kalliopi Benetos, Elena Cotos, Ann Devitt, and Antonette Shibani, 3–13. Berlin and New York: Springer.

Kruse, Otto, Christian Rapp, Chris M. Anson, Kalliopi Benetos, Elena Cotos, Ann Devitt, and Antonette Shibani (eds). 2023. *Digital Writing Technologies in Higher Education. Theory, Research, and Practice*. Springer.

Lardilleux, Adrien, Serge Fleury, and Georgeta Cislaru. 2013. "Allongos: Longitudinal Alignment for the Genetic Study of Writers' Drafts." *Lecture Notes in Computer Sciences* 7817: 537–548.

Leblay, Christophe, and Gilles Caporossi. 2015. "A Graph Theory Approach to Online Writing Data Visualization." In *Writing at the Crossroads: The Process-Product*

Interface, edited by Georgeta Cislaru, 171–182. Amsterdam–Philadelphia: John Benjamins.

Lebrave, Jean-Louis. 1987. "Le Jeu de l'énonciation en allemand d'après les variantes manuscrites des brouillons de H. Heine." Thèse de Doctorat d'Etat: Université Paris-Sorbonne.

Leijten, Mariëlle, and Luuk Van Waes. 2006. "Inputlog: New Perspectives on the Logging of On-Line Writing." In *Computer Keystroke Logging and Writing: Methods and Applications*, edited by Kirk P. H. Sullivan and Eva Lindgren, 73–94. Amsterdam: Elsevier.

Levy, C. Michael, and Sarah Ransdell. 1995. "Is Writing as Difficult as It Seems?" *Memory and Cognition* 23: 767–779.

Levy, C. Michael, and Sarah Ransdell. 1996. "Writing Signatures." In *The Science of Writing: Theories, Methods, Individual Differences and Applications*, edited by C. Michael Levy and Sarah E. Ransdell, 149–162. Mahwah, NJ: Lawrence Erlbaum.

Lindgren, E., and K. Sullivan (eds). 2019. *Observing Writing: Insights from Keystroke Logging and Handwriting*. Leiden/Boston: Brill.

McGillivray, B., T. Poibeau, and P. Ruiz Fabo. 2020. "Digital Humanities and Natural Language Processing: « Je t'aime... Moi non plus »." https://doi.org/10.17863.

Towards a Digital Corpus-Based Method for Assessing Language Level in EFL Student Writing: a case study on Romanian undergraduate literary analyses

Alexandru ORAVIȚAN
Mădălina CHITEZ

Introduction

In Romania, the issue of using digital tools and corpus-based methods to assess student writing remains largely unexplored. As interest is on the rise with regard to producing new assessment frameworks targeting writing among EFL (English as a foreign language) students in particular, the Common European Framework of Reference (CEFR) for languages constitutes a powerful authority, as it continues to be used as the main point of reference for evaluating student languages levels. Throughout Central and Eastern Europe, there is significant demand for assessment practice updates that integrate digital tools and that contribute to streamlining the evaluation process and to producing reliable CEFR-based scores. This chapter treads new ground by proposing a digital corpus-based method for assessing language levels in EFL student writing in the aggregate, while paying particular attention to the case of literary analyses written by students who study English in Romania.

Research on student writing in Eastern Europe, and Romania in particular, has intensified only over the last decade, and systematic inquiries into related regional and national data are scarce and still ongoing (Doroholschi *et al.* 2018, 6–7). One area where significant strides are being made is the humanities, since writing is an integral part of university curricula within such disciplines. In Romanian philology faculties, for example, student writing has been used as a tool with which students can practice their language skills, and instructors can assess student progress. At the undergraduate level, writing practice now primarily serves to create the conditions under which students can develop the independent research skills required to

complete the final BA paper and to pursue postgraduate studies. Perhaps nowhere are these skills more in-demand than in the case of literary analyses, which students produce during seminars or term examinations.

Crașovan and Rogobete (2020, 94) provide a valuable overview of how literary analyses can be used as tools for assessing Romanian students of philology. In this survey, they stress the flexibility of the genre in terms of probing and developing students' skills in understanding and interpreting literary texts. Based on a corpus of student texts written in Romanian (L1), the study reveals some of the main issues that arise when undergraduates produce literary analyses of Romanian texts in academic settings: the inaccurate contextualisation of the literary work under scrutiny, the misuse of terms, and a subservience to the marking scheme whereby the rubric is treated as the definitive structural model with which to comply rather than as a template that facilitates the development of more original critical thought (Crașovan and Rogobete 2020, 111).

Indeed, awareness of the specificities of the genre remains a crucial issue when attempting to implement pedagogical practices for academic writing in Romanian universities, especially in the case of L1 writing in Romanian. Arguably also playing a pivotal role in the context for achieving performance is the discipline, which may expose students to a rich range of writing practices while also helping them to develop familiarity with academic language. In the context of Romanian universities, much of this pedagogical effort has been left to the practical courses and modules offered by foreign language departments, which has led to a paradoxical situation: "it may be a greater challenge for the Romanian students to handle academic texts in their mother tongue than in a foreign language" (Chitez 2018, 135).

However, while Romanian students of foreign languages may benefit from more systematic and structured exposure to reading and producing academic writing, one issue that is decidedly connected to the quality of the texts submitted for assessment—but which has yet to be the object of substantial analytic scrutiny—is the level of linguistic competence demonstrated in students' L2 writing. It is arguably this dimension that achieves the connection between the receptive and productive halves of literary analysis, since analyzing a literary text in L2 implies having both a clear understanding of the set foreign-language text as well as the ability to express

an informed critical opinion in a foreign language in writing. This conjunction of the productive and the receptive may also be viewed as one of the main aims of foreign language study programs at Romanian universities: to foster among students the ability to build practical communication skills in the languages studied in the program, cultivate critical analysis skills for literary studies, and develop intercultural skills (Facultatea de Litere, Istorie și Teologie, UVT 2021).

Indeed, assessing language levels in the case of L2 productive and receptive skills has been a preoccupation in the Central and Eastern European educational context ever since the Common European Framework of Reference (CEFR) for languages—which was formally introduced in 2001—was piloted in the 1990^{s} (Little 2011, 2). This significant development led to fundamental changes in foreign language teaching across educational systems, with one of the ripple effects being the introduction of criteria for quality assessment and test design, and a decisive shift towards skills-based teaching (Bérešová 2011, 178–181). In the case of the Romanian university system, foreign language study programs have integrated the CEFR into their everyday practice, with notable efforts being made to impose a minimum required level of linguistic proficiency for BA graduates. For example, in the case of the Faculty of Letters, History and Theology at the West University of Timișoara (Romania), the reference level B2—which denotes an upper-intermediate independent user—is used as the minimum threshold for the BA graduation examination in both of the study programs that are conducted in foreign languages: Languages and Literatures, and Applied Modern Languages. One of the exam requirements for final-year students in the Languages and Literatures undergraduate program, for instance, is to produce an on-the-spot analysis of a literary text from the curricula (Facultatea de Litere, Istorie și Teologie, UVT 2021, 5).

Therefore, language level assessments of literary analyses produced by students writing in L2 should be a primary concern when it comes to developing pedagogical frameworks that are appropriate for foreign language programs and that seek to integrate advances in digital technology and corpus-based methodologies. Focused on literary analyses written in L2 English by undergraduate students, this paper draws on the Corpus of Romanian Academic Genres (entitled ROGER)—a bilingual corpus which consists of three years' worth of academic papers by L1 Romanian and

L2 English students (2018–2021) and which was compiled with the help of 27 collaborators from 9 Romanian universities. The paper draws on the ROGER corpus in attempt to provide a framework for assessing the CEFR language level of academic writing among L2 English students, by using digital tools to diagnose language level efficiently and to obtain an aggregate view of language level across the entire corpus. The working hypotheses of this case study are: (1) that the language level of L2 English students is more or less on par with that of L1 English students; (2) that digital tools and corpus-based methods can be used to gain valuable insights into more extensive studies on student writing pedagogy or on the washback effect on teaching and learning practices associated with student writing.

Methods

1. Assessing language levels

Born primarily out of a belief in a plurilingual Europe, the CEFR has—since its formal introduction in 2001 (Little 2011, 2)—become a legitimized permanent fixture in the assessment of linguistic proficiency both within and without the borders of the European Union, including when it comes to L2 language levels. Divided into three broad divisions (A: Basic User; B: Independent User; C: Proficient User) and six competence levels (A1: Breakthrough; A2: Waystage; B1: Threshold; B2: Vantage; C1: Effective Operational Proficiency; C2: Mastery), the CEFR enjoys broad international recognition and impacts language education on a global scale, particularly regarding evaluation and marking criteria. Its impact on assessment has occurred despite the fact that CEFR levels were not originally developed to rate learner performance (Zheng, Zhang and Yan 2016, 6). Moreover, with respect to how university language courses are designed and evaluated, the CEFR has sparked a series of debates, a pivotal one of which is connected to the minimum level of language proficiency a student is expected to attain upon completion of a given course (Little 2011, 11–13). Significant attention has been paid to this aspect, especially in the larger context of international language certification through standardized testing (such as Cambridge, TOEFL, IELTS etc.).

Over the past two decades, the goal of automating language proficiency assessment by aligning user-generated messages to the CEFR has been a constant

preoccupation for scholars both within and beyond the realm of digital humanities. Some of the research ventures in this area have had a considerable degree of success, both at the written and spoken levels (*see* Ballier *et al.* 2019, Wang *et al.* 2018, Schmalz and Brutti 2021). In the case of various digital tools that have been developed, the assessment is based on learner dictionaries that use the CEFR levels to target vocabulary relevant to specific learner categories. Some of these tools have been branded as self-assessment tools (see English Vocabulary Profile Online 2015, CEFR Checker 2021, Self-assess your language skills with our CEFR tool 2020) or employ the advantages of NLP (Bengoetxea, Gonzalez-Dios and Aguirregoitia, AzterTest: Open Source Linguistic and Stylistic Analysis Tool 2020). Using such tools to assess a corpus of student writing may offer the benefit of accessing swift generalizable information about language proficiency at the macro level.

2. Corpus

The corpus used for this study is a subset of the bilingual Corpus of Romanian Academic Genres (ROGER). The research project also dubbed ROGER, conducted at the West University of Timișoara (Romania), aimed to develop a corpus-based methodology to analyze academic writing genres at Romanian universities from a contrastive perspective, comparing genres written in Romanian and genres written in English (Bercuci and Chitez 2019, 737–738). The searchable ROGER corpus is part of the deliverables stemming from this project (Chitez, Bercuci *et al.* 2021). The subset used for this research venture is divided into two study corpora: the ROGER_LIT_EN corpus, which consists of literary analyses produced by BA-level students; and the ROGER_LIT_EN_BA corpus, which consists of BA papers on literary topics. ROGER_LIT_EN contains short literary analyses mostly about authors (such as William Shakespeare, Charles Dickens, and modernist writers like James Joyce and Virginia Woolf) who have become central to English-language literature syllabi; these texts have been produced within the setting of BA programs in English from three Romanian universities: the West University of Timișoara (65 texts), the "Alexandru Ioan Cuza" University of Iași (98 texts), and the University of Bucharest (2 texts). The ROGER_LIT_EN_BA corpus comprises BA papers on

literary topics written within the last five years at the West University of Timișoara (16 texts) and at the University of Bucharest (4 texts).

To gain a comparative understanding of L2-L1 dynamics, we have compared the two ROGER study corpora against an L1 reference subset, taken from the British Academic Written English Corpus (BAWE). This subset contains 93 L1 student texts that pertain to the discipline of English and that mainly consist of essays on and critiques of literary topics (Nesi *et al.* 2008).

Corpus	Number of texts	Word count	Average
ROGER_LIT_EN	165	263,942	1599
ROGER_LIT_EN_BA	20	366,210	18310
BAWE_English_L1	93	212,964	2290

Table 1: Corpora profiles

3. Tool

An online and open-source NLP tool, MultiAzterTest has been chosen as the most suitable digital tool for this study, because it automates the process of rating the English texts in digital format according to the CEFR, and because it makes it possible to establish the incidence of words pertaining to a specific language level. MultiAzterTest is capable of providing over 125 linguistic measures for texts in English, Spanish and Basque, but these can also be easily adapted to other languages (Bengoetxea and Gonzalez-Dios 2021). In its first iteration as AzterTest, the tool was designed to assess the readability of texts and outline their main stylistic features (Bengoetxea, Gonzalez-Dios and Aguirregoitia 2020). The newer multilingual version, however, can calculate scores for a range of features, such as lexical diversity, classical readability formulae, word frequencies, and morphology and syntax. The feature relevant to this research project is vocabulary knowledge, as processing an English text with MultiAzterTest yields the number and incidence of words at each CEFR level (Bengoetxea and Gonzalez-Dios 2021, 11). The wordlist that we used was taken from the Oxford Learners Dictionary, which consists of 3,000 core words, at the A1 through B2 levels, and an additional 2,000 words, at the B2

through C1 levels (Bengoetxea and Gonzalez-Dios 2021, 7). Following experimentation and analysis that used a sequential minimal optimisation (SMO) classifier, the MultiAzterTest obtained a 69.29% accuracy score for its vocabulary knowledge feature (Bengoetxea and Gonzalez-Dios 2021, 22).

Results

All the texts making up the corpora under study have been processed using the MultiAzterTest's descriptive and vocabulary knowledge modules. *Figure 1* outlines the results obtained (in absolute values), in terms of the number of words identified in each CEFR level.

It should be noted that the word count associated with the "not in A1-C1" category merely implies that these words are neither part of the 3,000 core A1-B2 nor part of the 2,000 supplementary B2-C1 wordlists provided by the Oxford Learners Dictionary; arguably, these words reflect specialized vocabulary or proper names discussed in the texts themselves or in the in-text citations.

A broader understanding of the language level present in the corpora can be gained from the incidence of vocabulary per 1,000 words classified according to the CEFR levels, as shown in *Figure 2*.

An interuniversity comparison according to the incidence per 1,000 words, contrasted with the BAWE subset, yields the results for ROGER_LIT_EN and ROGER_LIT_EN_BA presented in *Figure 3*.

Discussion and conclusion

Overall, both when it comes to word count (*Figure 1*) and incidence per 1,000 words (*Figure 2*), the hypothesis of the closeness between undergraduates' L2 English language levels and that of L1 students is partially confirmed by the results. These similarities are particularly evident at the intermediate and advanced levels, where the values scored by the ROGER study corpora are increasingly similar to those recorded for BAWE_English_L1. More specifically, the combined incidence per 1000 words for the B1-B2 thresholds is 59.02 for ROGER_LIT_EN, 64.58 for ROGER_LIT_EN_BA, and 71.04 for BAWE_English_L1, which situates the L2 BA papers in close proximity to the texts by L1 students. For level C1, the results are all

the more striking in that ROGER_LIT_EN_BA scored a higher incidence (16.99) than BAWE_English_L1 (16.1).

The results thus confirm the assumption that implementing CEFR-based pedagogy in the context of L2 writing among students will lead to success in language proficiency, particularly in the context of an existing language level threshold for final-year examinations. Moreover, while limited in scope, this study demonstrates that digital tools and corpus-based research methods can be used for the purposes of assessing language level not only in the case of undergraduate student writing but at all stages of developing EFL productive skills according to the CEFR. Thus, new grounds are established for imple-menting a new corpus-based method in the assessment of EFL-learner writing, and such implementation can be used to gain insight into the teaching and assessment of academic writing among learners, at the interuniversity level.

Annex 1: Figures

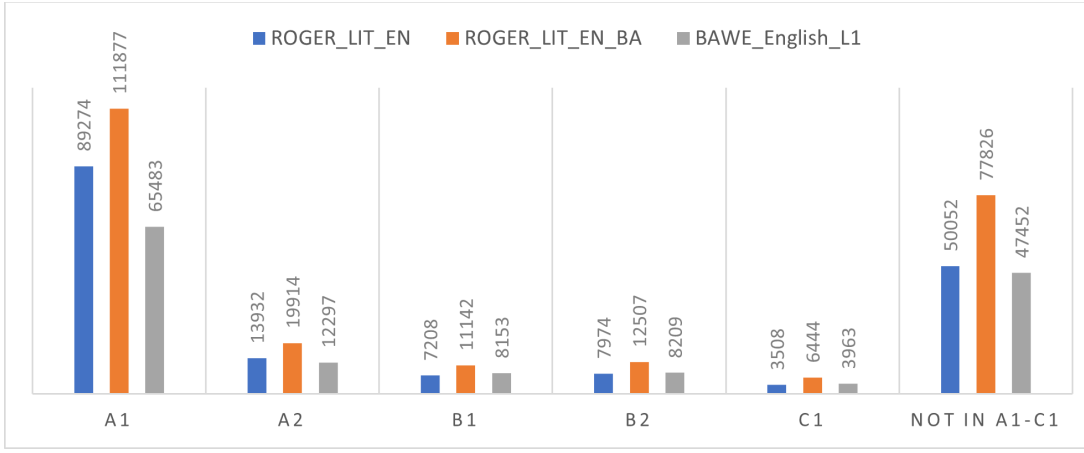


Figure 1: Word count - CEFR levels

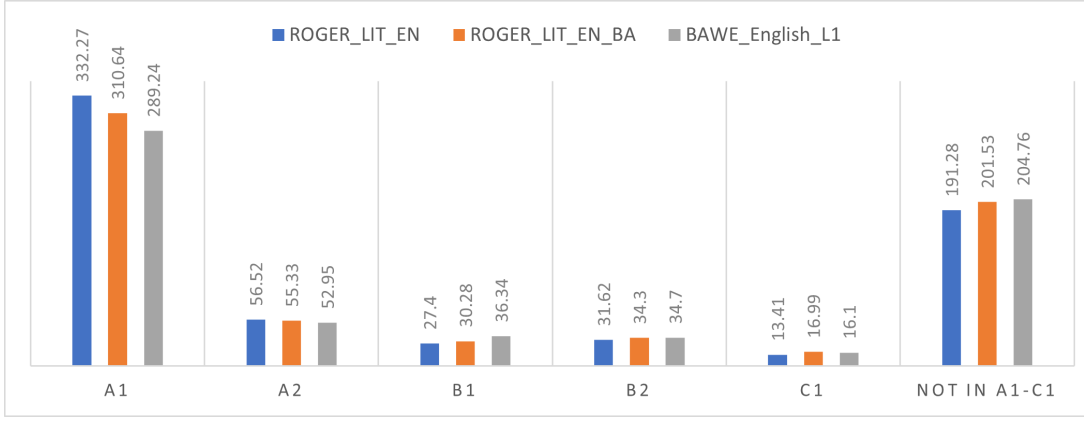


Figure 2: Incidence per 1k words

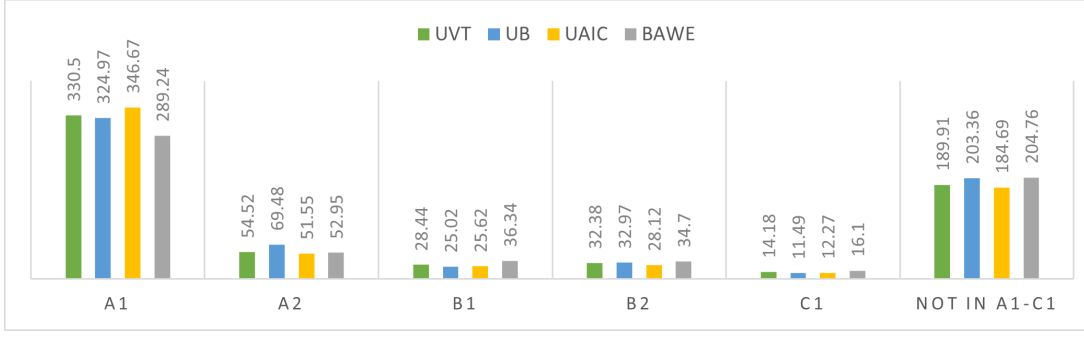


Figure 3: ROGER_LIT_EN Interuniversity incidence per 1k vs. BAWE_English_L1

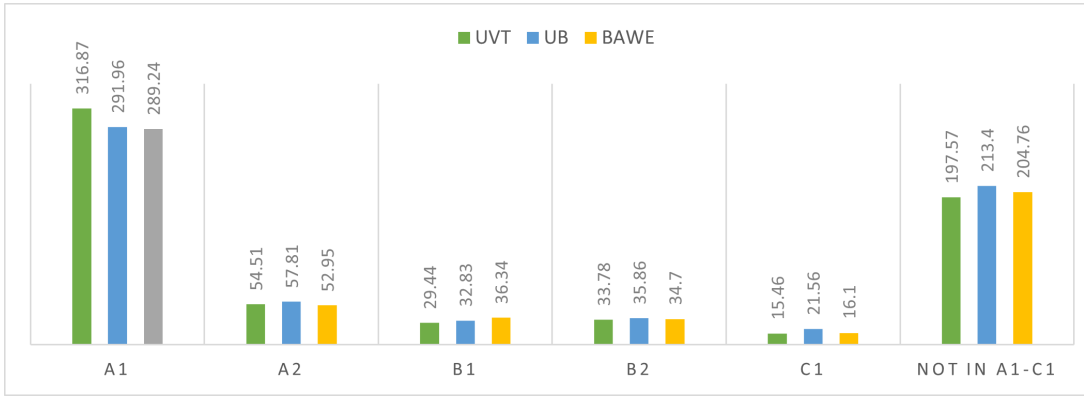


Figure 4: ROGER_LIT_EN_BA Interuniversity incidence per 1k vs. BAWE_English_L1

References

Ballier, Nicolas, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk. 2019. "A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors." In *Transforming Learning with Meaningful Technologies. EC-TEL 2019. Lecture Notes in Computer Science*, edited by Scheffel M, Broisin J, Pammer-Schindler V, Ioannou A and Schneider J. Cham: Springer. doi:10.1007/978-3-030-29736-7_23.

Bengoetxea, Kepa, and Itziar Gonzalez-Dios. 2021. "MultiAzterTest: a Multilingual Analyzer on Multiple Levels of Language for Readability Assessment." *ArXiv–Computation and Language* (Cornell University). Accessed February 10, 2022. https://arxiv.org/pdf/2109.04870.

Bengoetxea, Kepa, Itziar Gonzalez-Dios, and Amaia Aguirregoitia. 2020. "AzterTest: Open Source Linguistic and Stylistic Analysis Tool." *Procesamiento del Lenguaje Natural* (6): 61–68.

Bercuci, Loredana, and Mădălina Chitez. 2019. "A corpus analysis of argumentative structures in ESP writing." *International Online Journal of Education and Teaching (IOJET)* 6 (4): 733–747.

Bérešová, Jana. 2011. "The Impact of the Common European Framework of Reference on Teaching and Testing in Central and Eastern European Context." *Synergies Europe* (6): 177–190.

2021. *CEFR Checker.* Cathoven AI. Accessed February 10, 2022. https://www.cathoven.com/en/cefr-checker/.

Chitez, Mădălina. 2018. "Corpus Linguistics Meets Academic Writing: Examples of Applications in the Romanian EFL Context." In *University Writing in Central and Eastern Europe: Tradition, Transition, and Innovation*, edited by Mădălina Chitez, Claudia Ioana Doroholschi, Otto Kruse, Lukasz Salski and Dumitru Tucan, 133-148. Cham: Springer.

Chitez, Mădălina, Loredana Bercuci, Andreea Dincă, Roxana Rogobete, and Karla Csürös. 2021. *Corpus of Romanian Academic Genres (ROGER).* Centre for Corpus-Related Digital Approaches to Humanities. https://roger-corpus.org/.

Crașovan, Elena, and Roxana Rogobete. 2020. "Analiza de text literar: o perspectivă cantitativă asupra textelor studențești." *The Annals of the West University of Timișoara. Humanities Series* 58: 93–116.

Doroholschi, Claudia Ioana, Dumitru Tucan, Mădălina Chitez, and Otto Kruse. 2018. "Introduction: Understanding Academic Writing in the Context of Central and Eastern European Higher Education." In *University Writing in Central and Eastern Europe: Tradition, Transition, and Innovation*, edited by Mădălina Chitez, Claudia Ioana Doroholschi, Otto Kruse, Salski Łukasz and Dumitru Tucan, 1–14. Cham: Springer.

2015. *English Vocabulary Profile Online.* Cambridge University Press. Accessed February 10, 2022. https://www.englishprofile.org/wordlists/evp.

Facultatea de Litere, Istorie și Teologie, UVT. 2021. "Plan de învățământ anul universitar 2021-2022 valabil pentru anii II–III de studiu." Undergraduate Studies Programme - Curriculum, Department of Modern Languages and Literature, Faculty of Letters, History and Theology, West University of Timișoara, Timișoara. Accessed November 1, 2021. https://drive.google.com/file/d/10q_3NMzoGsOHvoTbItCXuwmnoxj-S2rwZ/view?usp=sharing.

Facultatea de Litere, Istorie și Teologie, UVT. 2021. "The Regulation for the Organizing and Carrying Out the Graduation Exams of the Bachelor and Master Degree Studies at the Faculty of Letters, History and Theology." Regulation, Faculty of Letters, History and Theology, West University of Timisoara, Timisoara. https://drive.google.com/drive/folders/1v7kfdsWa7Z1ik6mNbzzFuZNRh1yKsGkf.

Little, David. 2011. "The Common European Framework of Reference for Languages, the European Language Portfolio, and language learning in higher education." *CercleS* (DeGruyter Mouton) 1 (1): 1–21. doi:10.1515/cercles-2011-0001.

Nesi, Hilary, Sheena Gardner, Paul Thompson, and Paul Wickens. 2008. *British Academic Written English Corpus.* Oxford Text Archive. Accessed February 10, 2022. http://hdl.handle.net/20.500.12024/2539.

Schmalz, Veronica Juliana, and Alessio Brutti. 2021. "Automatic Assessment of English CEFR Levels Using BERT Embeddings." *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)* (CEUR Workshop Proceedings) 3033. http://ceur-ws.org/Vol-3033/paper14.pdf.

2020. *Self-assess your language skills with our CEFR tool.* 26 June. Accessed February 10, 2022. https://www.pipplet.com/self-assessment-scale.

Wang, Y, M.J.F Gales, K.M Knill, K Kyriakopoulos, A Malinin, R.C van Dalen, and M Rashid. 2018. "Towards automatic assessment of spontaneous spoken English." *Speech Communication* 104: 47-56. doi:10.1016/j.specom.2018.09.002.

Zheng, Ying, Yanyan Zhang, and Youyang Yan. 2016. "Investigating the practice of The Common European Framework of Reference for Languages (CEFR) outside Europe: a case study on the assessment of writing in English in China." *ELT Research Papers* (British Council) 16 (1): 1–20.

DIGITIZING ROMANIAN HERITAGE

A Computer-Assisted Analysis of Eminescu's *Doină*

Ioana GALLERON

In a paper published in 2023, Patraș, Galleron and Patraș pleaded for a digital scholarly edition of Mihai Eminescu's work, using *Doină* as a case study for defining preferred editorial principles and technical choices. As shown repeatedly over time, digital editions have the advantage of being accessible and interoperable; also, to the difference to the paper editions, the one provided for *Doină* appears easier to follow.[1] In addition to this rationale, the present article aims to consolidate the argument by providing examples of how such an endeavor might facilitate the interpretation of literary texts. Thanks to a large array of tools and methods, digital editions bring to the fore patterns that are unseen or difficult to perceive, thereby contributing to the necessary and constant reappraisal of the literary patrimony. Without claiming—in the name of a dubious "scientific approach"—to replace the experience of reading for pleasure, such methods enable the testing of intuitions and the discovery of unexpected features; in short, they enable an increase in the body of knowledge and the esthetic emotion that are gained through literary works of art.

Without going into detail about the choice of the poem, or about the technical and philological choices made for *Doină*, it is necessary to summarize the ideas

[1] When Nicolae Iorga writes about Perpessicius's work, which he views as exemplary, he explains: "De altfel, matematica adnotatorului va putea fi […] foarte cu greu urmarită. Eu unul n-o pot face ușor: uneori mă pierd. Era mai bine, orice proporții ar fi luat ediția, să se dea și mai des, cum se făgăduise, integralitatea versiunilor. 'Etaj', 'subsol', ca 'terminologie arhitectonică', îngreuie rezultatul acestei admirabile osteneli. Fără a mai vorbi de semne, a căror amintire nu poate urmări pe cititor" (*In addition, editor's mathematics of annotation will be… very difficult to follow. I cannot do it easily: sometimes I get lost. It would have been better, whatever the volume of the book, to offer more frequently, in accordance to what has been promised, the full versions. 'Main text', 'footnotes', all this 'architectural terminology' weighs down on this admirable endeavor. Not to mention the various signs, whose signification the reader may forget.*) (Iorga 1939–1940, 152-153).
[*All translations from Romanian to English belong to this chapter's author*].

developed in Patraş, Galleron and Patraş (2023) to gain a better understanding of the discussion that follows. Written between 1878[2] and 1882, *Doină* exists in 10 versions, which have been assigned the letters A through J.[3] They have been transcribed with a semi-diplomatic protocol, then encoded following the parallel segmentation method. The result can be visualized with Versioning Machine or any other tool supporting such annotation.[4] All verses from the ten versions (258 lines in total) have been compiled in a single TEI-xml document, presenting almost 800 variation points, whether phonetic, graphic, lexical, grammatical or otherwise. The sheer number of variants explains why, in print editions, scholars concentrate on so-called "significant variations"—that is, those that carry a high semantic load and offer an interesting point of view into the main choices made during the creative process. One of the main advantages of digital methods is to allow taking into account all variations, whether big or small. This leads to new insights into Eminescu's creative process of *Doină*. In what follows, an in-depth study of the thematic patterns, then of the melodic outline of the poem and finally of the

[2] The emergence of *Doină* seems to be have been precipitated by Eminescu's reaction to the decisions of the Congress of Berlin in June-July 1878, following the end of the war between The Russian Empire and The Ottoman Empire. Several amendments were made to the Treaty of San Stefano. In addition to granting Romanian independence—and even rewarding the country, by expanding its territory to include two thirds of Dobruja—the new Treaty of Berlin (as it is now known) also ratified the definitive loss of Bessarabia, much to Eminescu's wrath. Furthermore, the treaty required the Ottoman Empire, as well as the various newly formed nations, to recognize several civil rights for the Jews, which was a measure bitterly debated by the intellectual circles of those times (Eminescu included). Commenting the 'Jewish issue' as well as the poem's reference to *railways* ("*drum de fier*"), Călinescu contends that the starting point of the poem is the Strousberg scandal (Călinescu 1969(1932), 276). Indeed, in the aftermath of the Romanian War of Independence, the Romanian government negotiated with German shareholders in order to buy back the costly railway which had been built in part by the consortium directed by the Jewish entrepreneur Bethel Henry Strousberg. However, by the time Eminescu put the first verses of *Doină* on paper, this scandal was already quite old, which suggests that its contribution to the conception of Eminescu's text might be less important than Călinescu argues.

[3] As indicated in Patraș, Galleron and Patraș 2023, the designation of the versions follows Perpessicius for A and B. We distinguish two separate versions corresponding to Perpessicius's C; as a consequence, we propose 10 versions instead of 9, and the designating letters have been modified accordingly after C.

[4] The annotated file can be downloaded from https://nakala.fr/10.34847/nkl.ae92l4uz, DOI : [10.34847/nkl.ae92l4uz].

toponyms used in the text, will be conducted on the basis of these 800 variants, perused with digital data and methods.

Thematic patterns

In an interesting attempt to analyze *Doină* by adopting an interdisciplinary approach that, on the one hand, takes into account the belated diagnosis of Eminescu's bipolar disorder and, on the other hand, incorporates what the authors call "verbal information," Pîrvu *et al.* (2016) affirm that:

> the first part of the poem ("From Tisa to Dniester's tide/All Romanians to me cried,/That they could no longer dwell/Amidst the foreign swell... The foreigner is everywhere […]") may have been written also in 1882 (late summer, early autumn), whereas the second part ("He who loves the foes about/ […]/ Rise, O'Stephen, from the ground […]") will have been written during the hypomanic episode in late 1882 winter, early 1883 spring. The date for *Doina* [*sic*] should consequently be given as 1882–1883. (373)

While such a hypothesis is clearly ill-informed, ignoring the wealth of information uncovered by Perpessicius about the duration of the poem's creation process, Pîrvu *et al.* affirmation serves to draw the reader's attention to the rather multidimensional nature of the poem. *Doină* mixes several literary motifs that convey quite different messages. The use of the digital methods can help unravelling the various stages through which *Doină* has developed, from its first outline in 1878-1879 (version A) to the most recent known versions (I and J).

A count of the variants per version (annotated with the <rdg> tag in TEI) allows to establish both a rate of differentiation and a rate of similarity. The former is calculated as the number of cases in which a version brings a variant of any nature (spelling, grammar, addition, deletion, etc.) that does not appear in any of the other versions. From this point of view, the most original version of *Doină* would be F, with 63 own variant lines; the most "common ground" version turns out to be E, with only 33 own variant lines. It should be noted that the text published in *Convorbiri literare* (version J) is not one of the most "common ground" versions: with 56 own

variant lines, it is the second most distinct version. Most points of differentiation relate to spelling and punctuation, as Perpessicius observed; however, to the difference to Perpessicius, I consider these aspects to be not so neutral, since they contribute to revealing the dialectal Moldavian dimension of Eminescu's scriptural practice.

The similarity rate is more complex, requiring a pairwise calculation of versions. Given the exploratory nature of the present study, as well as the relatively limited number of variants, I calculated this rate based on XQuery extractions;[5] in the event of a larger-scale analysis (one involving a much larger number of poems and variants), the testing of other methods is necessary, such as a stemmatology program written in R,[6] or the automation of extractions with Python. For the present study, the collected values were entered into an Excel table with the different variants in double entry (where variants appear in both rows and columns), and a heat map has been generated (*Figure 1*[7]).

Several lessons can be learned from this visual representation. First, the heat map confirms that *Doină* evolved in a non-linear manner. Versions A and B are produced shortly after each other, and therefore share a great number of similarities. With versions C and D, Eminescu tries out what Perpessicius calls a new "pattern." It should be noted that version E—which constitutes a first stable version of the poem, written neatly and maybe intended for publication in the "Familia" journal—resembles version C more than it does version D; also, to some extent, version E is a return to the earliest attempts (A and B). Versions F and H explore themes not yet addressed in the previous versions; in between these two versions one finds version G, which largely returns to the thematic patterns of E. Version I revisits certain choices made in E and G, while also coinciding in many ways with A and B. Its

[5] Annotation as well as the XQuery extractions have been carried out using oXygen, https://www.oxygenxml.com/.

[6] For instance: https://cran.r-project.org/web//packages/stemmatology/vignettes/stemm-atology.pdf.

[7] The graph is to be read from left to right. To give but an example, the second value in the first line means that in 181 cases version B matches version A. The lowest number, 127, indicates that version H meets version F only in 127 variation points. Since each version necessarily matches itself, the meeting points between A and A, B and B, etc. are in deep green. The total number of variation points (that is, the total number of <app>) recorded in the TEI-xml file is 261.

colors in the graph tend to confirm Perpessicius's assumption, according to which this version I is the one Eminescu has read in Iași just before falling ill;[8] it can therefore be considered as a quasi-definitive version. However, it should be noted that the similarity rate of J with the previous versions (including I) is decreasing: contrary to what Perpessicius states, the orthographic differences are not trivial; rather, they distinguish the published version not only from the "copy from Iași" (if indeed this is what Eminescu shared with his listeners) but also from the other versions written by Eminescu.

With respect to the rate of similarity, it is worth emphasizing the special character of the H version. Perpessicius describes it as a last exercise before the definitive pattern. It is strange, however, to place this version after E and F, with which version H has less in common than with D, especially since the (rather imprecise) dating proposed by Perpessicius ultimately associates it with 1880 rather than with 1881–1882:

> For dating we are guided by a draft of verses 37–48 that appears in Ms. 2264, f° 378, on the back of a proof form, in the vicinity of and with the same writing as a note, for the newspaper, which heralds the decoration of Macedonski with *Bene merenti*, class I. (Perpessicius 1944, 16)

Alexandru Macedonski (1854–1920) having been decorated in 1880, the argument suggests that H was drafted before G. Therefore, either the fragment on folio 378 has nothing to do with version H and is instead linked to versions C and D, or the folio fragment is indeed associated with version H (in keeping with Perpessicius's claim), and H would therefore be considerably older than F and G, probably written before *Doină*'s first stable version E. All this leads to the idea of *Doină* having two great periods of creation. The first one precedes 1881–1882; the poet experiments with several thematic patterns, and expresses anger and a certain feeling of helplessness. The second period comes after 1881–1882, during which the

[8] According to Călinescu, Eminescu read *Doină* in front of the "Junimea" members only, and did not make it to read it the following day, during the gathering in front of the monument dedicated to the famous Moldavian king of the Middle Ages, Ștefan cel Mare [Stephen the Great] (see Călinescu 1969 (1932), 276 et sq.).

poem matures on the formal level, while the tone becomes more moderate and the overall orientation more pragmatic.

The alignment of the variants allows us to observe that, before 1881, Eminescu develops two literary motifs that are significantly reduced starting with version G. First, the poet delights in listing the people of various foreign nations: Tatars and Kalmyks, Gagauz people, Hungarians and even—in version F—Bulgarians and Greeks. Of these, only the Russians ("*muscalii*"), which were introduced in version C, remain in the later versions (I and J). The other nationalities are designated as "foreigners" or "enemies," nouns which preserve the polemical aspect of the text but reduce the contentious nature of the poem. Also, the elimination of the other demonyms allows the poem to focus on the Russians; after all, they were the main beneficiaries of the Berlin Treaty of 1878 through the annexation of Bessarabia, a main catalyst for the writing of *Doină*.

The two periods of creation are also clearly distinct with regards to the use of the curses. These are much more abundant in versions A, B, C and D. Appeals to interventions from the natural world, whose forces lead the homeland to cleanse itself of impure bodies, are present in all four of the versions written before 1881–1882. First, Eminescu references various diseases—such as the plague or the cholera; see also "partea boalei" (any form of disease) in H—that are likely to attack foreign swarms that attack helpless Romania. Second, he evokes natural catastrophes (blizzards, floods) that are capable of sweeping such swarms away, and he goes on to develop the image of a symbolic dismemberment by various animals, with dogs and crows mentioned in the final versions. These appeals to primitive—even chthonian (as of version C, the "desert" is invoked)—forces are all the more significant because the poem engages in a kind of alternation between them and invocations of Stephen the Great. The image of the king's resurrection is absent from versions C and D, which contain the largest number of curses. Under the influence of anger, Eminescu seems to adopt a form of magical thinking, according to which solutions to problems come from non-human forces. In this first period, the poem expresses feelings, rather than reflecting calls for action.

After 1881–1882, the profile of *Doină* changes. With the exception of version H, whose chronological location is debatable (as discussed above), the poem opts for

a decidedly apotheotic ending, calling for the return of Stephen the Great. The image of the return of the king appears from the very beginning of the creative process (i.e. in version A), but it wanes at certain moments, as we have already seen. Moreover, in versions A and B, Stephen the Great is primary the lead organizer of a campaign to scorch the ground in front of the invaders, rather than a revivalist figure. In the toned-down, more optimistic context of the later versions, the same literary motif of the king's return symbolizes Romanians' capacity to act of their own accord, by mobilizing their proven virtues.[9] Magical thinking, which calls on supernatural forces, does not completely disappear in the later versions of the poem (forests, animals, and even the core spirit of the nation continue to be invoked), but it no longer plays a central role. While nature seems to protect the suffering Romanians in the early versions, the situation is somewhat reversed in the later versions, in which it is nature that predominantly suffers and asks for protection. In this respect, the digital juxtaposition of the different versions shows that the development of the nature motif, which increases from version A to version F, is somewhat slower from version G to version J. Eminescu eliminates the allusion to the spruce leaves, and reduces the redundancy between "forests" (păduri) and "groves" (codri), present in versions C and D. A description of the negative impacts of the country's enemies thus replaces the evocation of an invasion ("the Tartar hordes are coming"), which brings to mind a medieval era and is therefore effective on the affective level but less cognitively pertinent for Romanian readers in the 1880s. Similarly, the poet develops the motif of a Romanian illness ("moving backwards as a crayfish," etc.): foreigners are not themselves harmful (as in the A to D variants) but are dangerous because of the effects they produce. This argument is meant to give an ethical basis to the xenophobic appeal.

To sum up, before 1881–1881 *Doină* expresses a terrible but also self-destructive reaction to the amputation of Romania of some of its territory. After 1881–1882, the poem offers rather a kind of diagnosis of the national problems, while outlining an action plan. We are referring here to an ethical and esthetic call to

[9] Stephen the Great ("Ștefane, Măria Ta") is not the only tutelary figure invoked: manuscript A shows that Eminescu also explored other historical or mythical characters, such as Matias Corvin and Romanian peasants (e.g. "Ioan", "Romanian brothers").

action, rather than to a political discourse, which is more developed in Eminescu's journalistic activity from the same period (Vatamaniuc 1996, Mocanu 2020). The poem is meant to revitalize the past and to reinvigorate the threatened nation, by reviving the ghost of Stephen the Great and the national desire for self-determination. This evolution may testify to a precocious depressive episode (to borrow the medical terminology used by Pîrvu *et al.*), which might have taken place in 1879–1880, followed by a manic episode in the second part of the creative period (1881–1882) and a subsequent low, which led to Eminescu's admission to the Caritas sanatorium (Neghina and Neghina 2011).

The melodic evolution of *Doină*

The preceding analysis of the thematic patterns in *Doină* calls into attention the evolution of the melodic outline of the poem, which seems to have interfered, at least in certain versions and to some extent, with the cognitive acceptance of the images created. The most conclusive evidence of this is to be found in version C, where Eminescu seems to focus on the musical dimension of the poem, sometimes to the detriment of the text's semantic dimension. In this version, one sees caterpillars eating "spruce flowers", and ravens that "carry flesh on thorns/ And bones through stalks." This is because the rhyming words for "caterpillars" (omizi) are not numerous, and because the verb "to leaf out" ("a înfrunzi"), which could have been an acceptable alternative, has already been used in verse 26. One of the few options left is to rhyme "omizi" with "molizi" (spruce), evoking in the process a little-known "spruce flower" (which, incidentally, seems not to be threatened by predators, at least when compared to pulp and needles). In version D, Eminescu replaces "flower" with "leaf", rendering the image more acceptable; also, he moves the flowering to the previous verse, in connection with elms and cornels (*cornus mas*[10]). The alliteration in "z" ("frunze" [leaves], "molizi" [spruce], "omizi" [caterpillars]) therefore constitutes the main driving force in the writing of these lines, somewhat to the

[10] The Latin name has been kept here because a satisfactory English-language translation of the Romanian word for "corn" has not been found.

detriment of the sense. After 1882, Eminescu focuses more on the image of depleting domestic resources, while the alliteration becomes less perceptible or disappears altogether. The poem's themes and melodic outlines come together, achieving those formal qualities of poetry that are easy to follow, pleasant to listen to, and familiar to most Romanian readers.

A larger study of the melodic outline of *Doină* confirms this conclusion. In order to conduct such research, a phonetic transcription of the 10 versions has been realized. For pragmatic reasons, this task has not been conducted in strict accordance with the recommendations of the international phonetic alphabet. While ChatGPT 3.5 (which was used to generate a first transcription) and the Phonemizer Python script[11] (which has been a second option) use several characters from this alphabet to indicate specificities of Romanian phonology (such as /t͡ʃ/ for the palatal sound present in "nici," or the short i sound (/‿i/) at the end of the same word), counting the annotated phonemes with Python proved to be more complicated, insofar as the scripts identify two elements (rather than one) in the first case (/t/ and /ʃ/) and add an extra grapheme /‿/ in the second case. Such errors can probably be corrected with the use of a more specialized dictionary; however, the aim here was not to achieve transcription accuracy, but to count the occurrences of phonetic characteristics playing a distinctive role in the evolution of the melodic outline of the poem. Such features include, among others, semi-closed vowels, semi-vowels (/j/ in particular), affricative and fricative consonants (especially /t͡ʃ/, /d͡z/ and /ʃ/). The decision was therefore made to produce a customized phonetic transcription, using graphemes such as /ğ/ (instead of /d͡z/), /j/ (instead of /i̯/), /ç/ (instead of /t͡ʃ/ or /č/), and /wa/ for the diphthong "oa." While such workarounds imply a certain amount of post-processing (e.g. to discriminate between the real "j"—as in "jidan"—and the /j/ indicating a semi-vowel), they allow the phonetic transcription to be carried out easily and efficiently. In total, 30 distinct phonemes were identified, during the transcription process, as playing a major role in the creation of the melodic outline of the poem, and were subsequently counted through a simple character tally in Python.

[11] https://bootphon.github.io/phonemizer/.

In order to understand the results of the phonetic count, we need to appreciate two additional aspects that must be taken into account. First, the lack of precision in Eminescu's use of diacritics had to be addressed. The transcription described in Patraş, Galleron and Patraş (2023) attempted to follow Eminescu's manuscript as closely as possible; diacritics have not been added if they were missing in the manuscript, nor the various signs corrected, that Eminescu used to indicate the phonetic value of a given phoneme. However, such scrupulous respect for the letter of the text was not possible within the framework of phonetic transcription; as Vladimir Streinu[12] has shown, the editor's reluctance to engage his or her responsibility with regards to the phonetic value of a glyph leads to a kind of falsification of Eminescu's poetry, creating nonexistent and non-aesthetic pronunciations. For the sake of consistency, all instances of "a" have been rendered as "ă" in the appropriate contexts (e.g. "sătmar" instead of "satmar"; "străinătate" instead of "străinatate" or "strainatate"), even if this meant erasing the difference between the versions in which the well-known verse "poor, poor country" is spelled in standard Romanian ("săracă țară săracă") and the versions that feature the spelling "saracă țară, saracă!", with a more pronounced Moldavian accent.

[12] "Atât de stăpânit de superstiția rimei s-a dovedit unul dintre [editori], […], că intervine în două feluri: fie crezând că acordă mai muzical cuvintele neacordate deplin de poet, fie că îngăduie forme curioase la un cuvânt, care s-ar găsi în manuscris, pentru a salva… sonorizarea cu celălalt cuvânt. În amândouă cazurile, se ajunge la modificări flagrante față de textul eminescian" [One of the [editors] proved to be so possessed by the superstition of rhyme, … that he intervened in two ways: on the one hand, through giving a more musical tune to the words not fully tuned by the poet, on the other hand through allowing curious word forms, which are supposedly to be found in the manuscript, in order to save… the voicing with the other word. In both cases, these are blatant changes compared to the Eminescu's text]. Streinu also adds: "poetul tindea în grafia lui, chiar cea mai îngrijită, la uniformizarea semnelor pentru ă, â, sau î și i; de cele mai multe ori, același semn al lui ă indică și pe â sau î, semn care adesea devine o linie și nu o dată un simplu punct. […] Dar aparența manuscriselor, din cauza uniformității către care tind acele câteva semne grafice, nu e însăși realitatea lor, iar cât despre rimă, e de presupus că Eminescu, în rafinamentul auzului său, cuprindea și asonanța cu posibilitățile ei moderne, mai subtile decât ale rimei clasice." [Even when he used his neatest handwriting, the poet tended to employ the same signs for *ă* and *â*, or *î* and *i*; most of the time, the same diacritic as for ă also indicates â or î, becoming frequently a line and even a simple dot... But the appearance of the manuscripts, with the uniformity towards which these few graphic signs tend, is not their true reality, and as for the rhyme, it can be assumed that Eminescu, in the refinement of his hearing, also considered assonance with its modern and more subtle possibilities than the classical rhyme] (Streinu 1939, 103-105).

The second aspect to note in order to understand the results of the phonetic count is that transcribed versions do not take into account text fragments that appear in certain versions of the poem but not in others. These cuts have an obvious impact on the total number of vowels, semi-vowels, diphthongs etc. counted, since some occurrences may be present in the deleted parts but are absent or very limited in the remaining lines. One may therefore consider that some phonetic characteristics observed and interpreted in what follows have been given outsized importance. However, one must bear in mind that Eminescu himself considered the lines not taken into account as being less important for he was trying to express. On the melodic side, although the poet experimented with the same phonemes, words and patterns throughout the creation of *Doină*, some of them had clearly been given greater prominence in the building of the final melodic outline.

The different versions contain between 824 (version E) and 1216 (version F) phonemes. An analysis of the weight[13] of each phoneme in the total number of lines phonetically transcribed indicates that some sounds are indeed particular to certain versions. As hypothesized, such phonemes include the semi-vowel /j/, the affricative /t͡ʃ/ and the diphthongs "oa" and "ea."[14] Other discriminants are "a", "i", "j", "m", "n", "s", "u" and "v," with the two nasals ("m" and "n") and the fricative "v" being the most unexpected in this list. The presence of these three sounds is quite limited in certain versions ("m" in version A, "n" in version D, "v" in version E), an aspect that has not been observed so far since, when listening to a poem, one has more difficulty detecting such "missing" phonemes than identifying over-represented sounds. On the whole, vowels, semi-vowels and diphthongs tend to be the main discriminants between versions, suggesting that Eminescu was especialy interested in experimenting with assonance, as opposed to alliteration, in *Doină*.

[13] The weight has been determined as follows. I divided each phoneme count by the total number of phonemes in the version to calculate its percentage of contribution to the whole. After determining an average for each phoneme, across all versions, and a standard deviation (calculated with Pearson's formula), I noted the position of each phoneme in the data series. No value is situated beyond 2.36 standard deviation points, with only 14 values beyond -2/ +2 standard deviation points. All this indicates a fairly narrow distribution of data, and no real outliers; at the same time, this distribution supports the idea that some phonetic features are quite distinct in certain versions.

[14] The semi-closed vowel "ă" and the affricative "ț" did not meet the expectations regarding their discriminant power.

How has this work been carried out, and what can be said about the final product in terms of its melodic outline? In order to answer these questions, we have prepared a table that shows the distribution of the phonetic features of each version, and we have normalized and fitted the table for a principal component analysis (PCA), using a SKLearn script.[15] The results can be seen in *Figure 2*, generated with Matplotlib.[16]

The different versions of *Doină* can be divided into two main groups, with C, D and H very distinct on the y-axis. Significant differentiation can also be observed on the x-axis, between F, on the one hand, and C and E, on the other. On the whole, it is worthwhile noting that the distribution does not follow an arc that moves from the earliest to the latest versions: as observed when we analyzed the thematic patterns, Eminescu seems to have experimented with several possibilities, abandoning and returning to some of them, before reaching a certain stability of the melodic outline. In this regard, it is significant that, on the graph, version A is quite separate from the other versions: whereas some of its phonetic traits serve as the foundation for most of the subsequent versions, significant divergences can be noted as early as in version B. The relative weight of phonemes across versions (see note 15) indicates that version B reflects an increase in "u" sounds and a significant reduction of "j" sounds. Version C takes a sharp turn from the melodic outline of B, especially through an increase in the number of /j/ semivowels; this number remains very high in version H and quite high in version D, especially considering that this feature is very limited in almost all of the other versions. Version D incorporates several "i" and "r" sounds, signaling an interest in increasing the musicality of the text ("i" being a vowel, and "r" being a liquid consonant); it is also in this version that Eminescu explores the alliteration of "z" sounds, as seen earlier. Another significant change in phonetic choices occurs with version E. Rather than introducing new features, the poet reduces the phonetic diversity; this can be better understood by looking at the heat map in *Figure 3*, based on absolute values for each phoneme.

[15] https://scikit-learn.org/stable/index.html.
[16] https://matplotlib.org/.

Figure 3 is somewhat misleading, since it suggests that later versions contain a larger number of phonetic specificities than the earlier ones and does not reflect all of the observations made in the previous paragraph. This misleading effect is predominantly related to the difference in size between the versions, with versions F through J increasing considerably in the total number of phonemes, as opposed to the absolute values from versions A to E; indeed, the size of the poems—which accounts for almost 50% of the total variance—is the first component retained in the PCA. This visualization nonetheless allows us to see the extent to which version E represents a decrease when phonetically compared with traits from version D (especially), and explains why version F marks a new development in the musical conception of the poem. In terms of relative values, version F does not contain much innovation (apart from a noticeable increase of "ţ" sounds), but the increase in the word count changes its melodic outline. Despite certain local variations, such as the increase in the relative number of "a" sounds in version G or the similar evolution of "l," "î" and "n" sounds in versions I and J, version F becomes the foundation for the melodic outline of subsequent versions.

Once again, version H is distinctive because it clearly differs, in both absolute and relative values, from versions F and G, on the one hand, and from versions I and J, on the other. Version H can be named "a study in /t͡ʃ/" (e.g. "cine ne-au adus muscalii," "cine ne-au adus jidanii," "cine ne-au adus pe greci"), interspersed with a considerable amount of "ea" sounds (as seen in the examples quoted above). With respect to the first feature, this version is among those that most resemble version A, an argument which supports the suspicion that this version was elaborated during the first, rather than the second, period of composition of *Doină*. The melodic outline of version H is very distinct from that of versions G or I, and from that of almost all the other versions; it thus testifies to a chaotic creative moment, during which Eminescu explores different patterns, both figurative and melodic, before arriving at a more harmonious form of poetry in the period leading up to the 1883 recital in Iaşi.

Overall, the phonetic evolution of *Doină* shifts from "signature" versions—characterized by a surplus of one or several phonemes—to non-outstanding versions, in which all of the traits discussed above (vowels, semi-vowels, palatal fricatives, etc.) are carefully combined to give the impression of a steady flow of the laments,

of the imprecations about the foreigners and of the appeal to patriotism. As he works on toning down the chauvinistic dimension of the poem, Eminescu endeavors to soften its melodic outline. This softening can be seen as a trick used to inhibit critical reflection among readers or listeners who were—or, unfortunately, are—already inclined to adhere to the often heinous nationalist discourse espoused by the poet.[17] Adopting a more charitable view, we might instead appreciate this phonetic softening as an indication of how the poet was distancing himself from the extremely xenophobic comments he had initially made; as already noticed by scholars, Eminescu was rather ambivalent in his positions: "for almost every violent anti-Semitic statement from one article, another less violent one, which reevaluates and contradicts the former, can be found in another article or, sometimes, in the same article" (Butaru 2012, 119).[18] At any rate, the digital exploration of *Doină*'s phonetic dimension draws attention to the interrelated roles and evolution of musicality and ideology.

Drawing a Romanian map

Among *Doină*'s most distinctive features are the place names used in the opening line: "Nistru" and "Tisa." Yet few readers are able to quote other pairs of toponyms identified in the poem, despite the fact that these other place names are much more numerous (e.g. Sătmar, Boian, Turnu, Dorohoi). In fact, much like with the literary motif of the nature, the evolution of the 10 versions reflects an increase in the number of place names identified, followed by a stagnation. *Doină* is not the first poem in which pairs of toponyms introduced by "from XX to YY" appear; this

[17] To a certain extent, the role that the melodic outline played in the success of nationalist discourse is suggested by Lucian Tion (2021), who comments on the xenophobic drive of a poem titled *Scrisoarea a treia* (The Third Epistle): "The Bulgar and the Greek filth have been joined by 'the crook-faced Gagauz' […], and, of course, by the omnipresent Jew, who figures relentlessly in Eminescu's political writings as the primordial threat to Romanians everywhere, however, nowhere as melodious as in this example of poetic racist slur" (Tion 2021, 76).

[18] Butaru's explanation for this consistently contradictory position is that Eminescu's hate concerns not the foreign national *per se* but the foreigner as a vector of liberalism, carrier of a form of universalism and non-manual work ill-suited to an agrarian country. For other considerations about Eminescu's antisemitism and chauvinism, see Chiorean (2023) and Jicu (2012).

syntactic structure had already been used in *Bogdan Dragoș* and in *La arme* (*Pick up the weapons*), which Perpessicius referred to as "collateral" poems to *Doină*. However, Eminescu's text is certainly the one in which the phrase has almost an obsessive presence, which raises the question of its significance. Does Eminescu choose place names at random—for their euphonic virtues, for instance? Or does he make conscious choices that are semantically loaded? The first pair of toponyms suggests a rather deliberate selection, since the two rivers represent frontiers that are not only natural but also symbolic: for Eminescu, they designate the breadth of the Romanian territory, despite part of it having been confiscated by the Russian Empire. However, it is less clear what message he had hoped to send by using the other toponyms, some of which are placed at the very heart—versus on the margins—of Romania (Aiud, Brașov, Boian).

In order to better understand the meaning conveyed by these toponyms, I have projected their coordinates onto a map. The results can be seen in *Figure 4*, where green lines are used to connect the different pairs.

Though imprecise, this projection allows us to observe that the lines joining the pairing points cross out Romanian territories that were under foreign administration during the creation of *Doină*. The poem therefore very subtly expresses a sentiment of unacceptable loss, while simultaneously reintegrating the places into the national space. Rather than indicating frontiers and bringing the idea of a "round Romania" (known as "România dodoloață") to the fore, *Doină* therefore strives to visualize an amputated country. The idea of a poetical crossing out of lost territories is partly supported by a similar appeal that Eminescu made in one of his articles, which was written during the same period as *Doină*'s creation, and in which he invites the reader to "draw[…] a line from Hotin on the Dniepr to Prut."[19] Further evidence is to be found in the fact that toponym pairs privilege those geographical landmarks that optimize coverage of the homeland territory. This is probably why phrases such as "from Brașov to Aiud" (or Banat/Arad) and "from Olt to Criș"—two of the shortest diagonals among those drawn by the poem on the imaginary map of the Romanian-

[19] "Trăgând o linie de la Hotin din Nistru până la Prut" (Eminescu *qtd. in* Pleșca 2014, 39). The author does not give the exact references of the quotation, vaguely referring to a paper published in the conservative journal *Timpul* in January or February 1878.

speaking countries—are eliminated in the later versions. Overall, the accumulation of geographical points, while no doubt also informed by euphonic concerns, helps to establish, more and more clearly and from a version to the next, the image of an amputated Romania, adding to the apocalyptic view of a country that loses its vitality because of the foreigners.

Conclusion

The analysis of the literary motifs, phonemes and toponyms, as they appear and disappear from one version to another, allows us to better understand the creation process of *Doină.* Two main periods of elaboration have been observed, one during which the poem functions especially as a curse, and a second one where it primarily functions as a patriotic appeal. Eminescu's vision is perpetually evolving, and the three aspects analyzed in the previous sections support Perpessicius's intuition that the published version of *Doină* is not the final one. Had the poet not been affected by the mental illness that ultimately led to his death, we would probably have had to annotate additional versions of the poem, and maybe compare them with a stable, if not definitive, version published by Eminescu himself.

Much remains to be done in terms of using digital tools to exploit *Doină* more fully. A coding of the variants as (ortho)graphic, lexical, topographic, and so on, as called upon in Patraş, Galleron and Patraş (2023), still remains to be done: such coding would most certainly help to refine our understanding of the literary motifs, ideology and melodic outline of the text. Moreover, an analysis of the quantitative evolution of each part of speech might support (or contradict) some of the observations made above. This paper is, however, a good example of what might be accomplished if we had at our disposal a complete digital collection of Eminescu's texts—or at least a digital collection of those texts that were published during his lifetime.

The present paper is not the first attempt to read Eminescu's poetry with digital tools. As compared to previous attempts to describe Eminescu's style from a quantitative point of view (Briciu, 2019) or to apply a sentiment analysis to his work (Lupea, 2021), the specificity of this article lies both in the decision to focus on a

single text and in the exploitation of a carefully curated digital scholarly edition. It argues for an approach to Eminescu's work, and the literary corpus more generally, that is anchored in a more qualitative analysis, to the detriment of a distant reading. The digital humanities developed, in the 2000s, with the goal of continuing, amplifying and simplifying traditional literary criticism and philology; in recent years, however, the discipline has strayed greatly from these initial promises, especially in the context of a rapid development of natural language processing and large language models. These recent achievements undoubtedly have much to offer the exploration of cultural heritage, in all of its forms. However, making sense of this heritage still requires human intervention in the preparation and exploration of data. The present contribution hopes to have been a reminder of this truth.

Annex 1: Figures

	A	B	C	D	E	F	G	H	I	J
A	261	181	162	159	187	146	165	148	172	167
B		261	151	154	187	150	171	141	173	166
C			261	177	183	129	155	150	154	150
D				261	161	145	167	160	166	159
E					261	161	182	150	180	172
F						261	162	127	154	145
G							261	168	180	171
H								261	154	147
I									261	200
J										261

Figure 1: Similarity between versions

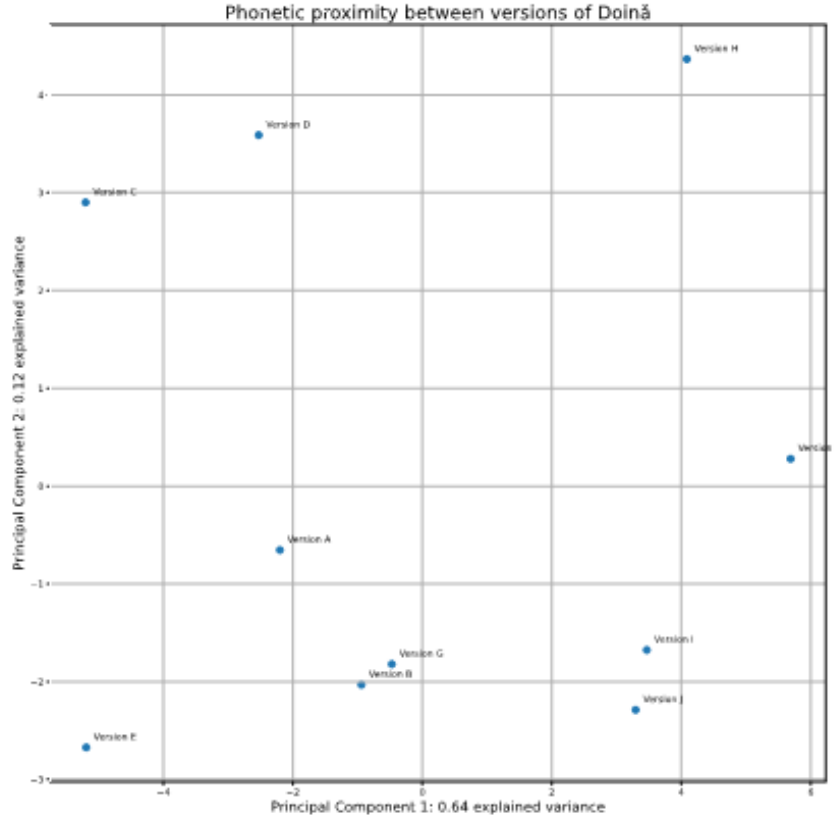


Figure 2: Principal component analysis of the phonetic transcriptions of versions of *Doină*

	Version A	Version B	Version C	Version D	Version E	Version F	Version G	Version H	Version I	Version J	Average
a	90	F	91	F	88	128	F	F	F	124	109,1
ă	F	F	45	F	47	65	F	F	F	F	53,5
b	F	F	F	F	4	F	F	12	F	F	7,1
c	F	25	F	48	23	F	F	56	F	F	37
ç	F	F	F	F	F	F	5	22	F	F	10
d	F	F	25	28	F	53	53	F	F	F	42,6
e	F	F	47	F	45	83	F	83	F	F	64,7
ea	7	F	F	7	F	F	F	24	F	F	12,9
f	F	3	F	7	2	8	3	F	F	F	4,9
ğ	F	F	F	F	F	F	F	F	F	0	1,4
g	F	F	2	4	F	F	F	F	9	10	6,6
h	2	F	3	3	F	3	F	7	F	7	4,8
î	F	F	17	17	F	F	F	F	31	31	23,4
i	F	F	F	74	48	72	55	71	F	F	63,2
/j/	37	F	F	F	33	53	F	58	53	F	45,8
j	4	F	F	F	F	F	F	F	F	F	1,2
l	F	F	34	F	34	52	F	F	51	51	41,5
m	26	F	28	F	29	47	F	46	F	F	37,1
n	F	F	68	69	65	F	F	103	101	F	84,8
o	F	F	27	F	F	46	F	32	F	F	39,6
p	F	F	F	F	F	29	F	27	F	F	22,8
r	61	F	F	F	56	F	F	F	81	F	71,3
ş	F	F	F	27	18	27	18	16	F	F	22,1
s	F	38	F	F	38	54	F	60	F	F	46,8
t	F	F	37	45	F	83	F	F	80	F	65,2
ţ	F	F	F	5	F	13	F	F	F	F	8,5
u	53	F	51	F	F	74	F	F	74	F	64,3
v	F	F	F	F	5	17	F	F	F	F	12
w	F	F	8	5	F	F	F	F	14	F	11,1
z	F	6	F	F	F	6	F	7	F	F	4,9
Total phonemes	947	989	834	942	824	1216	1014	1162	1143	1131	

Figure 3: Phoneme distribution in each version of *Doină*



Figure 4: Localization of toponyms in *Doină*

References

Briciu, Anamaria. 2019. "Quantitative Analysis of Style in Mihai Eminescu's Poetry." *Studia Univ. Babes Bolyai, Informatica* 64 (2). DOI: [10.24193/subbi.2019.2.06].

Butaru, Lucian T. 2012. "Eminescu and the Pattern of Romanian Antisemitism." *Studia UBB Europaea* 57 (3): 117–128.

Călinescu, George. 1969 [1932]. *Viața lui Mihai Eminescu*. In *Opere*, vol. 11. București: Editura pentru literatură.

Chiorean, Maria. 2023. "Racialized Modernity in Late-Nineteenth-Century Romanian Literature." *Metacritic Journal for Comparative Studies and Theory* 9 (1). DOI: [10.24193/mjcst.2023.15.02].

Iorga, Nicolae. 1981. "Eminescu în și din cea mai noua ediție." In N. Iorga, *Eminescu*, ediție îngrijită, studiu introductiv, note și bibliografie de Nicolae Liu, 150–178. Iași: Editura Junimea.

Iorga, Nicolae. 1981. "Note istorice asupra editarii operei poetice a lui M. Eminescu." In N. Iorga, *Eminescu*, ediție îngrijită, studiu introductiv, note și bibliografie de Nicolae Liu, 124–129. Iași: Editura Junimea.

Jicu, Adrian. 2012. "Antimodernii Eminescu și Caragiale." *Philologica Jassyensia* 8 (2 (16)): 65–75.

Lupea, Mihaela, Anamaria Briciu, și Elena Bostenaru. 2021. "Emotion-based Hierarchical Clustering of Romanian Poetry." *Studies in Informatics and Control* 30 (1): 109–118.

Mocanu, Mihaela. 2020. "Eminescu's Journalistic Works – Editing Approaches and Reading Patterns." *Philobiblon* 25 (1). DOI: [10.26424/philobib.2020.25.1.03].

Neghina, Raul and Adriana Maria Neghina. 2011. "Medical Controversies and Dilemmas in Discussions about the Illness and Death of Mihai Eminescu (1850–1889), Romania's National Poet." *Medical Problems of Performing Artists* 26 (1): 44–50.

Patraș, Roxana, Ioana Galleron, and Antonio Patraș. 2023. "Către o ediție digitală a operei lui Mihai Eminescu: alinierea in TEI-xml a versiunilor Doinei și vizualizarea cu Versioning Machine." *Philologica Jassyensia* 19 (1 (37)): 109–123.

Perpessicius. 1944. "[Note și comentarii la] *Doină*." In M. Eminescu, *Opere III. Poezii tipărite în timpul vieții. Note și variante: de la Doina la Kamadeva*, ediție critică îngrijită de Perpessicius. București: Fundația Regele Mihai I.

Pîrvu, Bogdan C., Mihai A. Stroe, and Rodica Nagy. 2016. "Towards a Linguistic Patho-Chronology: Dating Poetry Using a Language-and-Mind Approach (Case Study: Mihai Eminescu)." *Revue Roumaine de Linguistique* 61 (4): 361–377.

Plești, Nae-Simion. 2014. "Basarabia în Publicistica și Poezia lui Eminescu." *Philologia* 56 (mai-august): 39–43.

Streinu, Vladimir. 1989. "O ediție Eminescu." In V. Streinu, *Eminescu*, ediție îngrijită, prefață, note și indice de nume de Mihai Dragan, 95–115. Iași: Editura Junimea.

Tion, Lucian. 2021. “The Use of Language in the Making of Romanian Nationalism from Dimitrie Cantemir and Mihai Eminescu to the Iron Guard and National Communism.” *Acta Univ. Sapientiae, European and Regional Studies* 20: 63–92. DOI: [10.2478/auseur-2021-0012].

Vatamaniuc, Dimitrie. 1996. *Publicistica lui Eminescu, 1877–1883, 1888–1889*. București: Minerva.

Mapping Eastern European Political Discourse and Inequalities through Public Monuments: a digital cartography crowdsourcing project

Voica PUȘCAȘIU

Brief Introduction to Public Art Issues

This project combines a long-standing interest in art in public spaces, data-visualization tools in the Digital Humanities and, last but not least, an educational approach to addressing such matters. While art is never innocent, public art is particularly charged with political meaning and, when discussing it, one should be aware of its ideological discourses. Data-visualization may seem like a superficial mode to tackle these rather nuanced issues; what it manages to do, however, is to reveal several aspects at once. For instance, it establishes the frequency of certain kinds of representations and their distribution: Who is represented in public artworks, how often, and where are these representations found within the city? How important does local history appear to be, as compared to national history? Data-visualization also provides insights into inequalities: Who is left out? Where are the women and the other minorities? Who "owns" the city centre, and who is relegated to the peripheries? How does such representation compare across towns and regions in Romania, and between Romania and other Eastern European spaces? Considering that communities devoid of public memorials are extremely rare, I turned to crowdsourcing to try to answer some of these questions by creating and interactive digital map.

Before diving into the specifics of the project, we will clarify what constitutes public space, what the nature and function of public art are, and how these concepts are interconnected. Truly open public space is notoriously difficult to find (Dickenson 2021, 5), since there clearly are rules and regulations at play at all times; yet, for the purpose of this project, "public space" will be defined as a location that is in no way fenced off and that is free of charge and open to all members of the public at any time. The decision to define the notion in this way risks further perpetuating the myth of the

democratic public space, so in order to make a relevant point at a later stage in our discussion, we have included in the pilot map a location that falls outside the scope of this definition—a botanical garden that operates on a schedule and where entry is subject to an admission ticket.

As we will soon see, even in the public spaces that correspond more strictly to the definition above, all voices are not equally heard and not everyone is equally represented. This is problematic in itself since, if community is where most (collective) memory is acquired (Halbwachs as cited in Deegan 2018, 21), then representation within public spaces is a critical issue. This community-building narrative (Johnson 2002, 293) is also what drives most choices when it comes to art in public spaces, and it is here where the stakes are highest, since lack of representations in public art reinforces the overall lack of representation in the community which impacts minorities. As passers-by, we tend to take our surroundings for granted, as we do with most of our symbols, including maps, as we will see. Whether we realize it or not, however, there is a logic to what is presented in public spaces, and it can vary from overt propaganda to something much more subtle, though perhaps no less ideologically motivated. Art is never completely detached from the sociopolitical sphere and this is all the more true of public art (Levinson 2018, 32), since it essentially acts as a form of state speech (Valls 2019, 177). Moreover, the fact that it is commissioned for public consumption—and often paid for by taxes—does not necessarily mean that the artwork is intended to benefit the public (Hoffman 1991, 543).

One need only look at how the public responds to monuments to notice that the relationship between the viewers and the artworks is complicated and sometimes tense, although it can also be characterized by moments of tenderness. While these topics are for other studies to address in more detail, they are worth mentioning here because the public reception of statues lies at the heart of what initially motivated the present research.

Thus, we see how political realities, past and present, shape (art in) public spaces and are reflected in them. Yet who exactly is aware of these intricacies and is properly equipped to read and interpret public spaces in Romania? Surely someone who is to some degree a scholar of history could do it, but this seems less likely for most of the general public, and such analysis would definitely not be available to visitors who are unfamiliar with Romanian history. Since Romanian history, like the

Romanian language, occupies a more peripheral position on the world stage, what could a foreigner make of all the statues? Certain historical or literary figures who have been commemorated by public monuments need no introduction: for instance, many people worldwide know of Charles Dickens (1812-1870)—who, incidentally, has had surprisingly few memorials dedicated to him because he expressed a personal preference for such monuments to be limited in number—or Queen Victoria (1819-1901). The reasons why these figures are represented in public spaces are, much like the figures themselves, not in the least obscure. Yet such a claim cannot be made with the same level of conviction about Alexandru Ioan Cuza (1820-1873) or Ion Creangă (1837-1889), who are household names for Romanians but whose roles are largely unknown to people outside this culture. How could foreigners begin to understand the layers of power play and national discourse that are specific to a region when even the locals are often at a loss? Aside from the objects themselves and their placement, little to no explanation or context for the statues, monuments, plaques and memorials is provided, in Romanian or any other language; what's more, as we will see, much is lost in translation even when explanations are provided. Romania's case is by no means unique in its position or the complexities of its overtones, which is why we will also be comparing this case to that of other Eastern European countries which have certain historical backgrounds.

This brings us to the prime motivation of this entire initiative: the belief that public art is created not only to serve the lowest common denominator but rather to educate one and all. Our project of creating an interactive digital map of public monuments aims to do just that: to offer a broad overview of art in public spaces and to contextualize public art by using the ensuing database and visualization as a standing point for various types of analyses. The map can also be used as a popularization tool: while the monuments of one nation are arguably meant for its citizens, and the state should therefore not be concerned with what foreigners do or do not understand, the research of comparing and contrasting local and regional histories is not only legitimate but vital, since we all learn more about one culture when it is reflected in another, and we eventually gain a better understanding of the place that this culture occupies in broader terms, including continentally and internationally. Romania is still severely underrepresented (Patraș 2022, 133) in most (Eastern) European research groups and international repositories (Patraș 2022, 140), and this research contributes to closing such gaps.

Methodology for mapping and crowdsourcing

This project originated during the COVID-19 pandemic lockdown as an impromptu solution for students specializing in Art History at the Faculty of History and Philosophy at the Babeș-Bolyai University in Cluj-Napoca. The project is therefore a direct consequence of how COVID-19 affected academia, not only with respect to education and research but also insofar as falls under the category of everyday research practices in the (post-)pandemic era, as observed by Patraș, and it is still novel (Patraș 2022, 129).

In short, under normal circumstances, first- and second-year students at the bachelor level complete a summer internship which is compulsory and reflected in their credits. Students usually opt to carry out the internship in museums, galleries, or archives in either Cluj-Napoca or their hometown; during the pandemic, however, most cultural institutions were either closed altogether or not admitting entry to those not already affiliated with the organization, which meant that the usual internship options were unavailable. Since, by this time, classes had already been conducted online for the better part of the university year, everyone was tired of Zoom meetings and other online activities, I believed that documenting public monuments—a project conducted in outdoor spaces—would be a welcome change. It is also useful to remember that most students were back in their hometowns during this period, so this documentary activity could double as an opportunity for them to (re)discover familiar spaces, from the perspective of art history students.

Although public art has been the main focus of my own research, the topic has been absent from the art history curriculum at "Babeș-Bolyai" University and therefore represented a novel theme for the students. Initially, our goal was to gather and build a database of public statues in Romania, as no such database—and certainly nothing as complex and exhaustive as the one we were proposing—existed yet.[1] Since one of the students' tasks was to record the locations of the monuments, the idea of developing an interactive digital map was soon born. The workflow of the

[1] Since we began this project, an initiative of exhaustively documenting places of memory in Cluj-Napoca and the many forms they take has been published under the title of a database (Jakab). It is also possible that a digital version of this dataset will eventually be made available. Another project that documents and maps monuments in Romania and other nearby countries is *Monuments of Remembrance 1918–2018*, available at http://monuments-remembrance.eu/ro/.

data visualization itself was established at a later stage in the project, and it has continued to evolve as part of an ongoing process of introducing new generations of students to data visualisation notions and to methodologies related to not only to such visualization but also to the field of Digital Art History more generally.

Data collection

The initial task given to students was to identify three-dimensional monuments and statues that were independent structures (as opposed to being affixed to a building), to photograph the object, and to list several related pieces of information in a Google Sheet. While it was mandatory to fill in certain columns, completing others was optional and could be done at a later date. Choices about which data were most urgent were made in collaboration with the students, and it was intended to encourage them to focus on the medium of sculpture, rather than losing themselves the methodological task of distinguishing statues on buildings versus ornaments. Plaques were excluded from the project because transcription would be required and students felt that working with such quantities of text would involve a departure from the visual component of art history research. The mandatory columns were:

- the identifying number of the photograph (all photos were saved in a dedicated Google Drive folder);
- the title of the artwork;
- the name of the town or village where the artwork was located;
- the address associated with the statue (or the best approximation of the artwork's location);
- the medium of the artwork (e.g. stone, wood, cast metal);
- the theme of the artwork (e.g. historical, religious, allegorical);
- the artwork's type (e.g. equestrian, bust, full figure, non-representational);
- the number of figures featured in the artwork (if applicable);
- the gender of the figures featured in the artwork (if applicable).

Since this information was readily available simply just by observing the works, filling in the spreadsheet was not thought to represent a lot of additional work. The optional columns were:

- the artist's name;
- the year when the artwork was erected;

- a summary of the figure's biographical data (if applicable);
- the role the figure was their best-known role (if applicable).

Looking back, we appreciate that this spreadsheet is only an approximation of a proper monument chart, as it lacks emphasis on technical details; nonetheless, it was an appropriate tool for undergraduate students. The students carried out their tasks with various degrees of dedication. Some students took many photographs of the statues, from various angles; others took only one photo; in all cases, however, any inscriptions were legible. There was great variation in terms of how the medium category was defined, with students sometimes improvising names for the materials used. For example, cast metal was identified as "cast bronze," "bronze," "metal," "forged metal," "iron" and even "bronze?"; and stone was identified as any number of different types of stones, some of which were not plausible. These numerous and significant variations meant that in order for the database to be usable, considerable time and energy was needed to normalize the data.

The students worked on this data collection throughout summer 2020 and attended a final meeting—where they received their grades for the work—at the beginning of September of the same year. Each student was required to identify and collect data on a minimum of five monuments, but many students exceeded this quota and, especially when travelling to different towns, continued ad hoc to provide information on additional monuments even after the assignment had officially ended. Thus, to date, the database contains more than 250 monuments from 47 different towns in Romania. Although the distribution of the monuments is by no means uniform and many monument-rich towns are altogether missing from the survey, the database offers sufficient information to draw some preliminary conclusions and to start constructing a visualization.

Populating the map

Cartography and history have always been connected, as maps inform geographic and cultural identities (Thomas 2013, 336). However, since the digital turn and the widespread availability of Global Positioning System (GPS) technologies, digital maps have become objects of our time; we even navigate differently because of them (Wrisley 2014, 96). Digital cartography is a subdiscipline concerned with the employment of information technology in the creation and use of maps (Visvalingam 1990, 115). The first Geographic Information System (GIS) was

used as early as 1964 (Farman 2010, 870), and as software becomes more readily available, mapmaking is no longer the exclusive domain of geographers; indeed, digital humanists of all types have started wielding their metadata to create spatial visualizations. The user's status is also changing from mere readers towards a more active role as digital maps allow greater customization and interaction (Farman 2010, 869) in both populating and using the map.

What's more, since maps can be seen historically as tools of colonization (Thomas 2013, 336), there is an increased usage of the word "mapping" in the context of subversive activities. In such cases, there is sometimes no actual map involved, yet the word is used as a synonym for documenting, or rather pinpointing forgotten gestures and locations which are generally missing from official cartographies produced or approved by the establishment (Kelly and Vassell 2023) and the digital realm is an outlet for this new meaning of mapping.

Considering the tensions inherent in public spaces, and the role that public spaces play in maintaining national narratives, a spatial visualization of public monuments has the potential to reveal patterns that are otherwise difficult to observe. Eventually, one might see how trends fluctuate over time and which artistic styles are more of less dominant. Digital maps also allow for virtual visualization of monuments that have been destroyed or displaced, which adds another layer of understanding with respect to political preferences and regime as it tracks them over time. Believing that the students who had assembled the database deserved some kind of project closure and that such a coda would also serve as a teaching opportunity to address how to create a visualization based on a dataset, I set out to create a workflow that would be functional yet simple enough for students with limited prior experience in data visualization software to follow.

Thus was born the pilot map, which is available online.[2] In addition to representing an opportunity to work out methodological kinks, it acts as an incentive to continue mapping out monuments and to expand the database. For this pilot, I have limited the geographic range of the monuments to those located in the city of Cluj-Napoca. One factor underlying this decision is my own proximity to these monuments; another factor is that they had already been fairly thoroughly documented, especially in popularization publications (Ciorca 2022; Lukács 2005).

[2] Voica Pușcașiu. "Sculptură din spațiul public din Cluj-Napoca." Web Map available at https://arcg.is/1eu8eP.

What ultimately enabled me to create this map was the course entitled Introduction to Digital Humanities, which I completed through the StarUBB Institute (Babeș-Bolyai University), and which was taught by Dr. Rada Varga in the DigiHUBB Centre. Without this experience, the path would have been infinitely more difficult. Using the methods taught in this course, I could appreciate the interconnectedness of different software programs, which offered me greater freedom to move strategically between them. While not necessarily the most straightforward method, the approach was advantageous because of its adaptability and the fact that it makes use of software versions that are free of charge. I feel compelled to highlight this last point, since I was determined to make the project as easy and low-cost as possible for future students to use, and since I did not want to use Google My Maps—a tool which is very useful but has limited customization options.

Starting off with just under 50 monuments, I created the pilot map using two programs: QGIS and the free browser version of ArcGIS. First, each point that represents the location of a monument was individually plotted in QGIS, with Google Roads used as a base layer for clarity of the representation and overal familiarity of people with the Google Maps visuals. While time-consuming and somewhat tedious, this activity is simple enough once a rhythm has been established. The initial spreadsheet was not used, instead the data in it were cleaned up prior to input, but otherwise, the new spreadsheet column headings for each point basically matched those in the students' database:

- ID
- Title
- Material (bronze or stone)
- Theme (e.g. historical, genre)
- Type (e.g. statue, bust, equestrian)
- Chrt. Nr. (number of characters)
- M/F (for the gender of the figures, or neutral for non-representational works)
- Inscript. (the inscription)
- Photo (dedicated to the URL of the photograph)

All of the fields were string type, except for "Chrt. Nr." which was obviously an integer. The initial character limit was set at 200 for the inscriptions, and this proved sufficient for all but one of the 47 points.

The layer created in QGIS was then exported and uploaded to ArcGIS, this time with a more whimsical, doodle-like layer used as a basemap. At first, the legend took into account the type of monument, and a different color was therefore distributed to each category, resulting in a vibrantly colored map. Soon thereafter, however, I decided to feature the gender of the characters instead of the type of sculpture, and this led to fewer colors being used—red for women, blue for men, green for mixed, and gray for non-figurative works—but perhaps drove home a more important argument. The dots were also sized according to the number of the featured characters. This choice created a large gender-neutral green dot in the middle of the city belonging to the multicharactered relief on the Obelisk located in Cluj's oldest square, but other than that, there's hardly a noticeable difference between dot sizes. The columns corresponding with each pin in QGIS were used in ArcGIS to create the text for the Custom Attribute Display, meaning the pop-up box that appears when each pin is clicked or hovered over with a mouse. Since I was still working for and occasionally with the students, the decision was made to phrase the Custom Attribute Display in a way that was grammatically coherent in Romanian. Basically, the sentence provides a brief description of the statue by taking into account the previously established columns. Thus, for each pin, the following text is generated: The sculpture [*Title*] is a [Type: bust / an equestrian statue / etc.] made out of [Material: bronze / stone / etc.], with a [Theme: historical / religious / etc.] subject. The box also includes the transcript of any inscriptions found on the statue and a picture of the monument.

Call for contributions

Defined as small, straightforward tasks completed online, crowdsourcing is increasingly used in heritage and educational institutions (Bonacchi *et al.* 2019, 2) and is particularly useful for not only dealing with undocumented and undigitized material but also increasing citizens' engagement with their national heritage (Ridge 2013, 435-436). Sometimes called "citizen science" (Terras 2015), these activities currently engage hundreds or even thousands of worldwide participants (Rotman *et al.* 2014, 110), who tend to choose tasks based on how the activities correspond to their personal interests (Rotman *et al.* 2014, 114) or existing know-how (Bonacchi *et al.* 2019, 11). Sometimes, the involvement of these participants represent the only way for researchers to build or work with a database. For various reasons, including

time, funding, and geographical distance or accessibility, one could choose to rely on contributions from the public, such was the case of this project from the very beginning since it was based directly on the student's participation in documenting the monuments.

A better term for the first part of the project might be "nichesourcing": on the one hand, the work did not involve calling on the general public to send in photos; on the other hand, it not only engaged experts(-in-training) (de Boer *et al.* 2012) but also reflected a certain amount of coersion insofar as the students' grades depended on their participation in the project, which means that one cannot truly gauge the participants' interest, even when—as previously stated—certain students contributed much more than others and over longer periods of time.

Following the decision to expand the database to account for public art in other former Eastern Bloc countries, the project entered a proper crowdsourcing phase, which is still ongoing with varied results. Initially, participants are enthusiastic, but this attitude soon subsides (Rotman *et al.* 2014, 118). In other words, there is a significant gap between one's intention to participate and their actual participation (Rotman *et al.* 2014, 114), this shift is usually attributed to a lack of time or a perception that the tasks are too technologically complicated (Rotman *et al.* 2014, 118).

The call for *Crowdsourcing Eastern European Public Monuments*, usually accompanied by a short oral presentation, was issued on the occasion of three different Erasmus+ mobilities during the second semester of the 2022–2023 academic year. These oral presentations took place at the University of Novi Sad (Serbia), Sofia University St. Kliment Ohridski (Bulgaria), and the Ss. Cyril and Methodius University in Skopje (Macedonia). Another channel for the call was a social media collaboration with the Contested Histories initiative and EuroClio (European Association of History Educators). By far, the greatest number of responses came from Serbia, receiving data for approximately 70 monuments; otherwise, the response rate has been low to nonexistent.

Of course, not all crowdsourcing projects are a success, and participation is highly dependent on the design of the call (Nookelegref *et al.* 2014, 25) and on project publicity, with most submissions being received within the first month of the call being issued (Nookelegref *et al*. 2014, 29). This reflects the fact that, despite their intentions, people typically forget about or skip work that is not on their

immediate to-do list. Also, research suggests that a very small cohort of the total number of volunteers actually do the work: Nookelegref *et al*. noted that, in one case study, 75% of the work was done by 10 contributors, while the other 25% was carried out by the other 366 participants (Nookelegref *et al*. 2014, 30). This seems consistent with what we have observed with respect to our call. Thus far, there are few respondents, but those who have contributed data tend to have done so for a large number of monuments—often more than 10 of them—and their submissions were made in the first couple of weeks of learning about the call. Occasionally, participants submit additional contributions later on, if they have visited a different city. Contributors tend to ask if their submissions were suitable, but they do not seem concerned about whether the task was completed correctly. They also seem to appreciate receiving a thank-you note as well as feedback on the statues for which they have submitted data. Our intention is to identify participants who have contributed to the project, but they are not necessarily interested in having their name associated with it.

Preliminary results and analysis

From as early as the very first student contributions, some obvious conclusions could be drawn from the data. For instance, it was clear which figure was most commonly being featured in towns around Romania: Mihai Eminescu (1850–1889), the national poet. He was followed closely by poet George Coșbuc (1866–1918) and national hero Mihai Viteazul (1558–1601), with the canonical visual appearance of the latter being nearly identical in all representations. These results are consistent with the idea that local histories are especially prevalent in public art and that the diversity of figures represented across the country is therefore great. Historical figures such as those alredy mentioned and monuments that commemorate the December 1989 Anti-Communist Revolution are rare instances of representations that speak to the nation as a whole and therefore seem to belong in all of the former historical provinces, which underscores the idea of Romanian unity, both politically and—perhaps more importantly—linguistically.

This project was conceived not only to develop suitable assignments for an art history course but also to introduce students to the idea of working with data collection and analysis approaches that differed from those to which they may already have been accustomed and to draw their attention to the importance of

gathering data which is open, connected, and smart (Oomen and Aroyo 2011, 138). Since the project ultimately centered around creating and interacting with a map, it has also represented a great opportunity to examine mapmaking itself from a critical perspective—especially in the light of the 1980s cultural turn of maps, which has enabled maps to be read in feminist, postcolonial and postmodernist ways (Cosgrove 2008, 162)—and to take advantage of maps' storytelling properties (Killing 2018). Of all the different types of images, maps are perhaps most likely to be considered neutral and, therefore, true (Farman 2010, 874); as a result, one easily forgets that a map's features largely reflect decisions made by the cartographer (Wrisley 2014, 102). It is for this reason that the visualisation was adjusted to highlight the representation of female figures among monuments. What better way to engage with feminist discourse and with the notion that maps will show us what we want to see?

Unsurprisingly, the pilot map featured very few women protagonists. Indeed, in terms of monuments located in true public spaces, only one statue features a female protagonist: the plague column,[3] which is situated far from Cluj-Napoca's city centre and features the Virgin Mary at its top. In addition to showing what cartographers want to highlight, maps—like all other data visualisations, including statistics—also serve to conceal. Reading the empty spaces between the map's dots shows that public spaces reflect a clear underrepresentation of women. Here, we return to the methodological exception flagged earlier: one of the reasons the Botanical Garden was ultimately included in our study—despite the fact that it is not considered to be a true public space—is that the visualisation would reveal a unique cluster of statues of women (representing nymphs and allegories). Reading the map's empty spaces also shows that public spaces reflect a dire underrepresentation of minorities. This is particularly worrisome considering society's racist undertones and

[3] The plague column represents an interesting and unique case study. As a religious symbol, the statue is not common in Romanian Orthodox representations; also, because it was intended as a symbol of the Habsburg Catholic regime and was initially located in the vicinity of the main building of the "Babeș-Bolyai" University, it was removed by the communist authorities in the 1960s and salvaged by a parish located on the periphery of the city, which is where it is currently featured on the map. However, mere days before the writing of this text, the statue was moved back to its original location in a commendable gesture that completed the makeover of the streets surrounding the former Jesuit Baroque building complex. Even though it has been months since the announcement was made, I intentionally elected not to change the corresponding pin on the map in order to further emphasize that maps are products of their time and can therefore be biased and become obsolete if not updated.

the troubled relationship that Romanians still cultivate with ethnic Roma groups whose histories are rarely discussed.

People are drawn to the ludic, which some generously and positively call gamification, but others rather maliciously call infantilisation (Lammes 2015, 200). However, we would do well to recognize mobile applications as meaningful tools for contemporary education (Jambeck and Johnsen 2015, 2), especially when it comes to GPS and interacting with a map, contributing to one, learning from one. The ArcGIS software allows mobile apps to be created based on a web map layer that draws on preset templates. These apps can be viewed on computer web browsers but are optimized for use on mobile devices. Thanks to a template called Nearby which creates a proximity-based app—all previously plotted information is available in app form,[4] and the monuments and their various details can be viewed on an adjustable radius based on either the viewer's location or another address of their choosing. This is an illustration of how such a map can be used on the go and *in situ*, allowing users to familiarize themselves with the statues they encounter and to learn more about them.

Conclusions and further questions

Cultural nuances are notoriously difficult to explain to those outside the culture in question. In Cluj-Napoca, the monument dedicated to the Szekely Martyrs is situated in an out- of-the-way area that is difficult to access (it is basically found under the main road) and houses few other statues or memorials; yet, in other towns such as Târgu Mureș, the same monument is located more prominently and has a park associated with it. How might one explain the rationale for these differences? In contrast with the Roman Catholic artistic tradition, the Eastern Orthodox Church—through its interpretation of the Ten Commandments—believes that three-dimensional representations should not exist. As a result, religious figures are featured in very few statues in Romania or Eastern European region. Yet, one day in 2018, three large bronze monuments were unveiled in Cluj-Napoca: while one of them was the long-awaited and much-debated statue of the Greek-Catholic Cardinal

[4] Voica Puşcaşiu. "Public Sculpture in Cluj-Napoca." https://www.arcgis.com/apps/instant/nearby/index.html?appid=f0c86c7209ce49fd9759dea78cf57aad.

Iuliu Hossu (1885–1970), the other two statues—of the Orthodox bishops Bartolomeu Anania (1921–2011) and of Nicolae Ivan (1855–1936), respectively—had hardly been mentioned in the public arena prior to the unveiling, which made these statues rather surprising. How does one explain this uncharacteristic presence if not as a race for the upper hand in representation?

Perhaps one cannot or should not explain such things. Yet a case could be made for promoting greater transparency. It is clear that debates over statues are not a thing of the past or even just a brand belonging to totalitarian regimes; they are still taking place. What's more, older interventions are still left, and, when no explanations are provided in public spaces, confusion and division among audiences grows. The commemorative plaques on the House of Matthias Corvinus (1445–1490) offer another example in Cluj-Napoca that is particularly illustrative of this race. The historical building features two such plaques (*Figure 1*), which present two very different, even conflicting, ideologies. As it turns out, we can as viewers be wildly ignorant about narratives being told in parallel within a same shared public space. The first plaque, which is written in Hungarian and dated 1888, uses a convoluted 19th-century style to tell the story of the House and is as much a memorial to Franz Joseph I of Austria (1830–1916) as it is to Matthias Corvinus.[5] The second plaque is bilingual (Romanian and English) and suspiciously undated, although a coat of arms and the inclusion of the English language of all things suggests that the plaque might only date back to 1996 (Lukács 2005, 53) during the administration of a mayor who was explicitly Romanian nationalist. As we can see here, apps and maps that offer *lingua franca* translations and additional information about interventions in public domain would openly showcase parallel ideologies; they could also contribute to demystifying ethnic struggles and demonstrating that—just like other items in public spaces—statues, plaques and memorials are products of their time and should be viewed as such.

[5] The inscription on the plaque reads: "Matthias the Just, son of Hunyadi János and Szilágyi Erzsébet, was born in this house on the 27th of March 1445. The mighty king, out of grace, has released the house in which he was born from every obligation. George II Rákóczi, Prince of Transylvania, has reinforced this privilege. Our Apostolic King, Franz Joseph the 1st has honoured this house with his visit on the 23rd of September 1887, and with a gracious donation made possible that it (the house) would be marked for eternal memory. This plaque is a herald of grace and reverence, placed by the owner (of the house), the free royal city of Kolozsvár (now Cluj-Napoca), onto the birthhouse of its greatest son. 1888."

Finally, one of the greatest challenges in this project, which is also its greatest strength, is its multidisciplinarity. The work touches on interconnected domains such as art history, sociology of the arts, memory studies, and public policy and brings them all together in what is ultimately a product and resource typical for the digital humanities. Until recently, however, it has been difficult in Romania to secure funding; since scholars tend to prefer working alone (Patraș 2022, 136) and within their academic disciplines, it has also been challenging to gather a much-needed research group willing to collaborate. As information continues to be added to the database, following ongoing crowdsourcing calls, the tasks of organizing, mapping and analyzing the dataset therefore seem increasingly daunting. There is still hope, as citizen-based science projects are currently being encouraged through a platform that will soon be launched by "Babeș-Bolyai" University. However, it remains to be seen how heritage and cultural-based projects will fare in that space and whether there is an interest for this particular initiative.

Annex 1: Figures



Figure 1: Two samples of plaques in Cluj-Napoca

References

Bonacchi, Chiara *et al*. 2019. "Participation in Heritage Crowdsourcing." *Museum Management and Curatorship* 34 (2): 166–182.

Ciorca, Ioan. 2020. *Monumentele şi statuile Clujului*. Cluj-Napoca: Casa Cărții de Știință.

Contested Histories. Accessed November 2023. Available at https://contestedhistories.org/.

Cosgrove, Denis. 2008. "Cultural Cartography: Maps & Mapping in Cultural Geography." *Annales de Géographie* 660-661 (2–3): 159–178.

de Boer, Victor *et al.* 2012. "Nichesourcing: Harnessing the Power of Crowds of Experts." In Annette ten Teije *et al.*, *Knowledge Engineering and Knowledge Management. EKAW 2012*. Lecture Notes in Computer Science, vol. 7603, pp. 16–20.

Deegan, Connor. 2018. "Why do public monuments play such an important role in memory wars?" *Constellations* 9 (1): 20–33.

Dickenson, Christopher P. 2021. *Public statues across time and cultures*. London and New York: Routledge.

EuroClio. Accessed November 2023. Available at https://euroclio.eu/.

Farman, Jason. 2010. "Mapping the digital empire: Google Earth and the Process of Postmodern Cartography." *New Media Society* 12 (6): 869–888.

Hoffman, Barbara. 1991. "Law for art's sake in the public realm." *Critical Inquiry* 17 (3): 540–573.

Jakab, Albert Zsolt. 2022. "Piatra aceasta a fost așezată… Locuri ale memoriei din Cluj (1440–2022)." Bază de date. Editura Institutului pentru Studierea Problemelor Minorităţilor Naţionale.

Jambeck, Jenna R., and Kyle Johnsen. 2015. "Marine Debris Tracker: Citizen-Based Litter and Marine Debris Data Collection and Mapping." *Computing in Science & Engineering* 17 (4): 20–26.

Johnson, Nuala C. 2002. "Mapping Monuments: The Shaping of Public Space and Cultural Identities." *Visual Communication* 1 (3): 293–298.

Kelly, Natasha A., and Olive Vassell (eds). 2023. *Mapping Black Europe: Monuments, Markers, Memories*. Bielefeld: Transcript.

Killing, Addison. 2018. "Building Digital Stories, Architecture and Cartography Meet Documentary and Journalism." *Architectural Design* 88 (5): 30–37.

Lammes, Sybille. 2015. "Digital Cartographies as Playful Practices." In *Playful Identities: The Ludification of Digital Media Cultures*, edited by Valerie Frissen. Amsterdam University Press, pp. 199–210.

Levinson, Sanford. 2018. *Written in stone: Public monuments in changing societies*. Durham: Duke University Press.

Lukács, József. 2005. *Povestea "oraşului-comoară": scurtă istorie a Clujului şi a monumentelor sale*. Cluj-Napoca: Biblioteca Apostrof.

Monmanier, Mark. 2018. *How to Lie with Maps*. 3rd ed. Chicago: University of Chicago Press.

Monuments of Remembrance 1918–2018. Accessed November 2023. Available at http://monuments-remembrance.eu/ro/.

Nookelegref, Julia, Angela Bartholomew, and Alexandra Eveleigh. 2014. "Modeling Crowdsourcing for Cultural Heritage." In Nancy Proctor and Richard Cherry (Eds.), *Museums and the Web 2014: Selected Papers from an International Conference*. Museums and the Web LLC, pp. 25–37.

Oomen, John, and Lora Aroyo. 2011. "Crowdsourcing in the Cultural Heritage Domain: Opportunities and Challenges." In *Proceedings of the 5th International Conference on Communities and Technologies*, pp. 138–149.

Patraș, Roxana. 2022. "Digital Romania: A Novelty Effect?" *Hermeneia* (29): 127–148.

Pușcașiu, Voica. "Sculptură din spațiul public din Cluj-Napoca." Accessed November 2023. Available at https://arcg.is/1eu8eP.

Pușcașiu, Voica. "Public Sculpture in Cluj-Napoca." Accessed November 2023. Available at https://www.arcgis.com/apps/instant/nearby/index.html?appid=f0c86c-7209ce49fd9759dea78cf57aad.

Ridge, Mia. 2013. "From Tagging to Theorizing: Deepening Engagement with Cultural Heritage through Crowdsourcing." *Curator: The Museum Journal* 56 (4): 435–450.

Rotman, Dana *et al*. 2014. "Motivations Affecting Initial and Long-Term Participation in Citizen Science Projects in Three Countries." In *IConference 2014 Proceedings*, pp. 110–124.

Terras, Melissa. 2015. "Crowdsourcing in Digital Humanities." In *A New Companion to Digital Humanities*, edited by Susan Schreibman, Ray Siemens, and John Unsworth. New Jersey: Wiley-Blackwell, pp. 420–438.

Thomas, Leah. 2013. "Cartographic and Literary Intersections: Digital Literary Cartographies, Digital Humanities, and Libraries and Archives." *Journal of Map and Geography Libraries: Advances in Geospatial Information Collections & Archives* 9 (3): 335–349.

Valls, Andrew. 2019. "What should become of Confederate monuments? A normative framework." *Public Affairs Quarterly* 33 (3): 177–194.

Visvalingam, Mahes. 1990. "Trends and Concerns in Digital Cartography." *Computer-Aided Design* 22 (3): 115–130.

Wrisley, David Joseph. 2014. "Spatial Humanities: An Agenda for Pre-modern Research". *Porphyra* 22 (XI): 96–107.

T

The Title of the Feuilleton Novel: a model of semantic annotation[1]

Lucreția PASCARIU

Introduction

This chapter aims to show how the titles of nineteenth-century feuilleton novels are constructed. By applying research methods tested in the field of computational literary studies, I have extracted and analyzed several entities/categories (*person names, place names, their determinants, role and gender mark*) from the feuilletons' titles, following the annotation model proposed by Patras *et al.* (2021). The semantic annotation, data processing and interpretation were devised in order to identify specific titling practices of Romanian feuilleton novels published between 1850 and 1920. With this approach, I was able to notice some titling *patterns*—for example, the shortening of titles—but also frequent formulas, "magic" formulas, we could say, since they were meant to attract readers. Because I wanted to test a methodology already applied in a previous project and to propose, in turn, an original approach—that is, comparing the results *of a semantic analysis* based on annotation with the results *of a sociological analysis,* based on a *reader survey*—I focused on the 66 titles of feuilleton novels proposed for the *POP-LITE* corpus.[2]

First, I reasserted the theories of paratext, which are considered by Genette to be a threshold and not a limit:

> For us, accordingly, the paratext is what enables a text to become a book and to be offered as such to its readers and, more generally, to the public. More than a boundary or a sealed border, the paratext is, rather, a threshold, or—a word Borges used apropos of a preface—a "vestibule" that offers the world at large the possibility of either stepping inside or turning back. It is an "undefined zone" between the inside and the

[1] This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS –UEFISCDI, project number PN-III-P1-1.1-TE-2019-0127, within PNCDI III.

[2] https://poplite.org/despre/, accessed 28 February 2023.

> outside, a zone without any hard and fast boundary on either the inward side (turned toward the text) or the outward side (turned toward the world's discourse about the text), an edge, or, as Philippe Lejeune put it, "a fringe of the printed text which in reality controls one's whole reading of the text" (Genette 1997, 3).

Of all the forms of paratext, the title in particular, presents itself in a problematic situation, since, however short it may be, it must fulfil, in relation to the text, *functions* such as *identification or designation, description, connotation.* If we turn to the consumer novel, the *function of temptation/ seduction*, the one that urges the reader to purchase a book or the serialized episodes of a story. As there is no data on circulation or sales, it is hard to imagine how the semi-literate Romanian public[3] reacted to the market stimuli. Therefore, my idea was to use the reactions of current audiences to the titles of the feuilletons in order to see which words in the title (called "entities" in my annotation scheme, *see below*) are perceived as more important or as essential for making a literary work desirable. The *sociological* dimension of my approach thus complements a distant-reading method, i.e. processing the data resulting from semantic annotation. If *the textual patterns* observed in the titling practices of the feuilleton novel reflect the tastes and, by implication, the stimuli embedded by nineteenth-century writers, then we should ask ourselves if twenty-first-century readers are still sensitive to these strategies.

Combining *distant reading*, *reader response* and *empirical literary studies* (Schmidt 1982), my investigation is meant to be extended to all the titles of Romanian feuilleton novels published throughout the nineteenth century (*cf. Dicționarul cronologic al romanului românesc*/ DCRR 2023). In the present research, the results of the analysis of the semantic units (*entities*) in 66 titles highlight the specificity of the feuilleton novel not only in terms of the complexity of the narrative formula, but also in terms of the names chosen to designate, connotate, and describe it.

Methodology

The corpus proposed by the project team *Pop-Lite: Romanian Popular Novels and Their Sub-genres during the Long Nineteenth Century: digital editing and corpus*

[3] There are studies that emphasize the illiteracy of masses during the long nineteenth century (Patras 2018, 29–33).

analysis comprises 120 novels (Patras 2023), illustrating, through representative and carefully proportioned samples, the main subgenres (historical novels, hajduk novels, mystery novels, sentimental novels, sensation novels, crime novels), periods marked by exponential productivity (about 650 novels published between 1850–1920, with large increases in number between decades).[4] Project members noted that of the approximately 650 novels published between 1850 and 1920, only 350 were published in volume and 279 in paperback form. Trying to replicate the proportion reflected in the DCRR, they aimed to include in the POP-LITE corpus samples from at least 1/3 of the feuilleton novels. The team thus included 66 novels published as journal instalments, representing 55% of the entire corpus, which indicates both the prevalence of this narrative form and its close connection with the phenomenon of consumer literature (Patras 2022).

After checking the list of proposed feuilleton novels for the POP-LITE corpus (data extracted from the DCRR), I went through the data in order to check the accuracy of the information by following the steps explained bellow:

a. Titles (with subtitles) were restored and the inaccuracies in the DCRR were corrected;

b. The annotation model of the title entities was applied and data reflecting possible correlations were extracted;

c. An empirical study was carried out through a title-reaction questionnaire distributed to 185 respondents; given the prospective methodology, the margin of error was not calculated;

d. The results obtained from the textual approach (annotation) and the empirical approach (questionnaire) were compared.

a. Restoring titles: By *restoration* I understand a process of checking and replacing titles enlisted in DCRR with original titles. Whether we want to notice a *pattern* in the titles or to uncover some kind of evolution, the first editions and their "title" page should be the reader's primary reference. Before applying the annotation model mentioned in the introduction, I noted differences between the titles recorded in the DCRR and the original titles of the feuilletons published throughout the nineteenth century. As a result, the 66 titles of the feuilleton novels included in the POP-LITE corpus have been checked with both DCCR entries and the original publication, i.e. periodicals and volumes. It

[4] According to *Dicționarul cronologic al romanului românesc* – DCRR (2023).

followed then that the rigorous restoration of the titles according to the original publications was mandatory.

Comparing the *Dictionary...* entries with the original titles, I arrived at two restoration strategies: *restoration by text-addition*—for titles not recorded in the DCRR according to the first edition/publication, and *restoration by replacement*, for titles where the DCRR recorded variants different from the original title (two or three distinct words or punctuation marks). It should be noted at the outset that I have not taken into account the changes occurred between the publication of the journal edition of the feuilleton (e.g. *Getta. Roman [Getta. Novel]*; *Copilașii Mărioarei. Roman [Mărioara's Little Children. Novel]*) and the edition of the same feuilleton republished in volume (e.g. *Getta. Roman* ***senzațional*** *[Getta.* ***Sensational*** *Novel]*; *Copilașii Marioarei. Roman* ***senzațional*** *[Mărioara's Little Children.* ***Sensational*** *Novel]*).

An example of title restoration by text-addition would be *Un tron pentru iubire. Mare roman de actualitate senzațională [A Throne for Love. Great Novel of Sensational Actuality]*, a novel published in 1910 in the newspaper "Minerva" by Const. Fulger (Vasile Pop). In the original publication, the place of the action and the historical moment are mentioned from the beginning: ***Revoluția din Portugalia sau*** *Un tron pentru iubire. Mare roman de actualitate senzațională* [***The Portuguese Revolution or*** *A Throne for Love. Great novel of Sensational Actuality*]. The restoration of the title was also applied to *Condamnata* [*The Female Convict*] by Emanoil Arghiropol, a novel published in Iasi in 1868 in "Jurnalul pentru toți". The title recorded in the DCRR is incomplete and the dictionary does not give any information on the subtitle: *Condamnata.* ***Tăcere în viaţă, iertare în ora morții*** [*The Female Convict.* ***Silence in Life, Forgiveness in the Hour of Death***]. Of course, there are also instances where genre labels (e.g. "novel", "novella") or sub-genre labels (e.g. "historical", "thriller", "mystery") are omitted from the DCRR, along with explanatory items indicating the presence of engravings (e.g. "illustrated with four engravings").

The restoration of the title by replacing elements was also necessary for the novel signed by Cloroform (Grigore Mărunțeanu), *Sfinx sau amintirile unei nopți tainice* [*Sphinx or the Memories of a Mysterious Night*], published in "Revista ilustrată" in 1899. The original version of the title does not refer to "the memories of a mysterious night", but to "the **secrets** of a mysterious night". In this case, I have

also applied the restoration by addition, the original title being entirely descriptive: *Sfinx sau* ***din secretele unei*** *nopți tainice.* ***Roman misterios în trei capitole întunecate*** [*Sphinx or* ***from the Secrets of a*** *Mysterious Night.* ***Mystery Novel in Three Dark Chapters***]. For punctuation, I recorded two cases of restoration by replacement, only one of which can be related to the action of the novel. *Măritată. Roman nou* [*Married. New Novel*] by Eugen Vaian was published in "Săptămâna ilustrată" in 1893. The action begins with the despair of a young woman, who realizes on her wedding night that her innocence and freedom are lost forever. This fact was graphically nuanced right from the novel's original title: *Măritată**!** Roman nou* [*Married**!** New novel*] [5].

2. Application of an annotation model: Three decades after the publication of Gerard Genette's landmark volume *Seuils...* researchers questioned the titling practices by applying a detailed semantic analysis to a multilingual corpus (ELTeC) including novels published between 1840 and 1920. The functions proposed by Gérard Genette, in particular the identification function and the descriptive function (Genette 1997, 80–94), were operationalized by defining a set of "entities". The connotative and seductive functions also remain defining, but they can only be checked by an in-depth investigation of relations between the annotated entities or by figuring out a way to measure the reactions of the readers at the titles' first sight.

The present chapter proposes a semantic analysis applied strictly to the titles of feuilleton novels published in Romania during the nineteenth century and included in the corpus Pop-Lite. In order to do so, I drew inspiration from the study *Thresholds to the "Great Unread": Titling Practices in Eleven ELTeC Collections* which interweaves computational literary studies with a chronological analysis of titling practices in various national and transnational contexts. Since the corpus providing the resource for the aforementioned analysis was built upon precise sampling criteria,[6] the comparison of titles in the 11 subcollections (English, French, German, Italian, Polish, Portuguese, Romanian, Serbian, Slovenian, Spanish, and Ukrainian)

[5] A full list of titles restored and used in the semantic annotation can be found in the appendices.

[6] The selection of texts involved the application of strict criteria such as: *chronology* (period 1840–1859 code T1; period 1860–1879: code T2; period 1880-1899: code T3; period 1900–1920: code T4); *gender representativeness* (between 10% and 50% female authors, 50% male authors, rest: anonymous authors/more than one author); *size* (short, medium and long novel: according to word count); *canonicity* (determined by number of reprints; high vs. low).

was carried out by applying a manual annotation scheme, focusing on the nouns and on their determiners. The authors' reasons for restricting the analysis to noun phrases in titles seem to be more complex, but two of them appear obvious: on the one hand, most titles are composed of noun phrases; on the other hand, the analysis of the most productive entities (that is, noun phrases) can highlight the contiguity of titles and subtitles (Patras 2021, 167). For this exploratory stage, I removed the syntactic component from the original annotation model, i.e. the hierarchical relations between entities, marked in the initial annotation schema with "center", "adjunct", "apposition", "determinant", "modifier".

But what do I mean by "entity"? In short, "entities" are (groups of) words that directly indicate persons, places, genres, (sub)genres of the novel or other elements of the title (such as *abstract entities* like "pe urma" [*following/ in the trail*], *temporal entities* like "războiului" [*the war*]) that may or may not stimulate the audience's imagination even before they actually start reading the text. The annotation of these entities was done manually by applying specific labels to each word. In a spreadsheet document, I entered the list of restored titles in the first column, then listed the entities in the following columns. Being the most productive category, the entities designating people allowed for a more complex semantic annotation. Thus, I was able to mark, beside the character's name ("Miul", "Mircea", "Nicu", "Ursuza"), his/ her status ("haiducul" [*the hajduk*], "cel bătrân" [*the old man*], "ciobănelul" [*the little shepherd*], "domnișoara" [*the young lady*]), and gender (male, female, not specified). For cases where the feuilleton titles did not contain a *personEntity* expressed by a proper noun, the common noun was labelled as both person and status: e.g. "condamnata" [*the convict*], "santinela" [*the sentinel*], "procuror" [*prosecutor*]. The remaining entities (places, abstract component elements, gender) were marked by copying the word—e.g. "Pitești", "moarte" [*death*], "gravuri" [*engravings*]—and confirming its presence in the title (yes/ no). I also marked how the title was constructed, e.g. by joining a title and a subtitle (yes/ no). The table below shows the extracted entities, labels used, description and some examples:

The entity	Tags/ Tag	Description	Example
place	yes, no	Identifiable geographical entities	America de Sud, Grivița
placeEntity	no, “noun”	Yes, copying the entity	Grivița
person	yes, no	Entity designated by either a proper noun or a common noun	Mircea, fiică
personStatus	no, “noun”	Yes, copying the entity	Mircea cel Bătrân
personEntity	no, “noun”	Yes, copying the entity	Mircea
gender	m, f, no	Gender specification	m
other	Yes, no	Entities that cannot be integrated into any of the above categories	Între artă și iubire
otherEntity	yes, “noun”	Yes, copying the entity	artă iubire
twoContentElement	yes, no	Titles with two elements: title and subtitle	Prințul Dimitrie. Roman
contentElementConnection	interpunc, word, mixed, no	How they are connected	interpunc
genreIndicator	yes, no	Reference to the (sub)gender present in the title	Roman
indicatorEntity	no, “noun”	Yes, copying the entity	Copilașii Mărioarei. Roman senzational

3. An empirical study: Based on the list of restored titles, we created an online questionnaire in which we asked respondents to note in the answer box the word in each title that caught their attention. The questionnaire was given to five age groups (10–14 years – 12%, 15–20 years – 20%, 21–25 years – 37%, 26–35 years – 23%, over 35 years – 8%) with different backgrounds (secondary school students, university students and teachers/ instructors). Data collection was carried out in two

stages: initial number of respondents: 104, final number: 185.[7] After correcting and standardizing the data (typing errors, use of non-Romanian specific characters, lowercasing of proper nouns, etc.) after each stage, we found that a higher number of respondents did not alter the initial "picture" obtained by querying approximately 100 people, therefore for some of the cases discussed below, I provide graphs from the first stage. The research question we started from is as clear as can be: What kind of entities will the twenty-first-century reader be drawn to first and foremost?

Profiling to the Romanian feuilleton

The semantic annotation of the 66 feuilleton titles led to the following data on the overall distribution of the number of entities. It seems that feuilleton authors prefer to integrate between one and seven entities in their titles. Titles with more than three entities are often descriptive, while titles with just one entity present an eponymous character, a plot space, an object or abstract/temporal component elements. However, from the overall distribution chart, titles with two or three entities caught my attention as they are also the most frequent (*Figure 1*: Overall distribution of entities).

After tagging all titles in the list, I also calculated the total of entities:

placeEntity—identifiable geographical entities, *personEntity*—person entities designated either by proper noun or by common noun, and *otherEntity*— sub(gender) indications and entities that cannot be integrated into any of the two categories. Out of 123 entities, the category *otherEntity* consists of 63 entities (51%), *personEntity* of 48 entities (38%), and *placeEntity* of 14 entities (11%). In what follows, I will analyze each category's traits and provide some examples.

otherEntity: By looking at samples of *otherEntities*, we could say that the feuilletons not only use words of/for impact, but also point to the reference reality, to the elements most familiar to nineteenth-century readers. The category *otherEntity* could also be associated with what Gérard Genette coins as "the temptation function",[8] as the words integrated in this category, vague and less familiar (therefore

[7] The survey data was collected during February – June 2023.

[8] "The fourth, of questionable efficacy, is the so-called temptation function. When present, it doubtless depends more on the third function (connotation) than on the second (description). When absent, too, I may add. Let us say, then, that the temptation function is always present but may

mysterious, to a certain extent), can create a horizon of cognitive tension. The distribution of non-personal and non-local entities (*otherEntities*, in my annotation schema) according to the novelists' gender needs further investigation. For example, negative-valence entities such as *war, jealousy, divorce, famine*, *enmity* or *catastrophe* are often chosen by male authors, while female novelists prefer entities with positive valences such as *love, life, reward* or *art*.

It is interesting how the twenty-first-century reader relates to these titles—some of them carrying about them an obsolete sound. As mentioned above, I asked respondents to indicate the word that most appeals to them in each of the 66 titles. As in any sociological analysis, errors are bound to occur. So, for titles consisting of a single word (4–5 cases of this kind) I considered the answers as null. The core question was *what mattered to them the most: was it the place of action, the character, the status or the other entities?*

The contemporary reader does not seem impressed with the category *otherEntities* when they are in the company of more "powerful" entities, for example the *personEntity*. For the novel *Condamnata. Tăcere în viaţă, iertare în ora morţii* [*The Female Convict. Silence in Life, Forgiveness in the Hour of Death*], the respondents showed a preference for the entity that denotes the character (*personEntity*), while the other entities (*silence*, *life*, *forgiveness*, *hour*, *death*)—many and varied, as can easily be seen—were chosen rather sparingly. Only one person, in the 15–20 age group, associated the person with the entities "life" and "death", considering *otherEntity* as being related to the character. By centralizing these data, what becomes noticeable is a hierarchy of importance that can be established within the *otherEntity* category (*forgiveness-death-silence* vs. *hour-life*). The words "death", "forgiveness" and "silence" (probably related to the character mentioned in the title) were chosen more times than "hour" and "life".

If the title contains a vaguer *personEntity* (*e.g.* a family status), the readers shift their focus to words with negative, mysterious valence, or those that have a strong impact, specific to the *OtherEntity* category. One such case is the anonymous novel *Drama din Constanţa sau Taina unei mame. Roman realist, petrecut în zilele noastre [Drama from Constanta or A Mother's Mystery. Realistic Novel Set in the Present Day*]. This time, the majority of readers are drawn to the word that excites (*drama*),

prove to be positive, negative, or nil depending on the receiver, who does not always conform to the sender's own idea of his addressee" (*cf.* Gérard Genette 1997, 93).

while the *personEntity* (*mother*) comes second in their preferences. Probably reflecting the readers' level of maturity and life experience, the 35+ age group considers the "mystery" to be more representative than the "drama". Even though the age group represents only 8% of the total respondents, it highlights the way the present-day reader perceives non-personal and non-local entities. The same is true for the 10–14 age group, which accounts for 12% of respondents. They tend to perceive the nominal group (*nowadays*) equally important with the *personEntity* (*mother*).

However, the reader's reaction to the *otherEntity* category changes when only abstract entities such as *love*, *following*, *love-making*, *hidden*, *art* appear. Here, the contemporary reader finds it difficult to choose a keyword. For example, across all age groups, there are people who have not been able to dissociate *otherEntities* from the titles of novels *Între artă și iubire. Roman* [*Between Art and Love. Novel*] by Sarina Cassvan and *Amor ascuns. Roman original* [*Hidden Love. Original Novel*] by M. Mihăescu-Nigrim.

personEntity: In spite of the list's relatively small length, I remarked a wealthy distribution of character mentions in titles, gender representation, i.e. occupations, status, origin or role in social relations. In the annotated titles, no more than two *personEntities* appear. Of the 66 titles, only 7 record the presence of the second character, who often relates to the first (e.g. ***Estella și Maria*** *sau Răsplata virtutei* [***Estella and Mary*** *or The Reward of Virtue*], ***Fiica*** *lui* ***Sejan***. *Poveste istorică* [***Sejan**'s* ***Daughter***. *Historical Tale*], ***Logodnica*** *principelui* ***Matei Basarab*** [***The Fiancée of*** *Prince* ***Matei Basarab***], ***Fata*** *lui* ***Hagi Dumitru*** [*The* ***Daughter*** *of* ***Hagi Dumitru***]), while 34 have only one character. Regarding the gender distribution of the *personEntity* category, I discovered an almost balanced distribution between male and female mentions for the first character (21/20), while for the second character there is an imbalance, with only two titles containing a secondary female character and five titles a secondary male character (2:5). These data are not relevant as such but they might turn relevant in a broader framework, as they open up the problematic of female and male social statuses that are used as anchors in the instalments' titles.

Probably the most difficult question is why female characters in the titles make up only 46% of all person-defining entities and not more, given the prejudices

regarding the female dominant of the feuilleton audience? Considering the prejudice that female readers (usually precariously educated women) are usually targeted as consumers of genre fiction, this is significantly lower than expected. And yet the relative balance between mentions of male and female characters might open a more nuanced perspective on both the gender of Romanian feuilleton consumers and the addressability of newspaper content. Scholars have shown the fact that the press and its current events were usually addressed to men, while fiction, inspired by inner life and fantasy, was concerned with the experiences of private, intimate space (Lyons 2003, 313–344). For Romania, however, we must also take into account the broad literacy process that began with the spread of mass consumer publications, as well as the introduction of mandatory education, which changed the dynamic of the feuilleton-consuming public (Anghelache *et al.* 2018, 27–32). In addition to its sensational subject matter and *reader-friendly* format, the feuilleton's popularity can also be explained by its role in educating the masses, as the novel—especially one that can be read in instalments—encourages the reader not only to identify with the character, but also to absorb the novel's socio-cultural references and intertexts. This could also explain the presence of such a varied number of *personEntities* in the annotated titles.

Often, the individuation of these *personEntities* is done by association with the occupations, status, origin or social role of the characters. The established categories refer to: *a. occupation*, when characters are defined by remunerated work: e.g. "procuror", "cioban", "santinelă" [*prosecutor*, *shepherd*, *sentinel*]; *b.* status, when characters are defined by position in the social hierarchy: e.g. "haiducul", "prințul", "sclava", "studentul", "voievozii", "orfani", "condamnata", "cel bătrân" (Mircea cel Bătrân), "cel frumos" (Radu cel Frumos) [*the hajduk*, *the prince*, *the (female) slave*, *the student*, *the voivodes*, *the orphans*, *the (female) convict*, *the elder* (Mircea the Elder), *the handsome* (Radu the Handsome)]; *c. origin,* when characters are defined by the social and cultural background e.g. "cazacul", "olteanca" [*the Cossack*, *the female from Olt county*]; *d. social role, when* characters play a role that has relational connotations, i.e. they are related to other individuals: e.g. "fiica", "sora", "copilașii" [*the daughter*, *the sister*, *the little children*]. More than half of the *personEntities* (36 out of 48) are associated with an occupation, status, origin or role. Of these, the best represented is the status category (*hajduk*, *prince*, *princess*, *married*, *young lady*, *fiancée*, *female slave*, *student*, *voivodes*, *orphans*, *convict*, *the elder/ Mircea the*

Elder, *the handsome/ Radu the Handsome*), which prompts us to reflect seriously on the sociology of the feuilleton and the way in which certain class or social role patterns are already fixed from the title.

Respondents who participated in the *reader survey* experiment confirm the observation that the character's social status "sells" the novel. They chose predominantly the status of the character, whereas character names were chosen only for titles recording a single *personEntity* represented by a proper noun (e.g. ***Elmira***. *Roman original în trei părți* [***Elmira***. *Original Novel in Three Parts*]). For example, when it came to the novels *Prințul Dimitrie. Roman* [*Prince Dimitrie. Novel*], *Căpitanul Ropotă* [*Captain Ropotă*] or *Miul haiducul. O pagină din epoca fanarioților. Nuvelă originală* [*Miul, the Hajduk. A Page from the Phanariot Era. Original Novella*] respondents seemed undecided about the key word of the title at first, but most of them noted in the answer box the status of the characters (*prince*, *captain*, *hajduk*) instead of the entity naming the person (*Dimitrie*, *Ropotă*, *Miul*). For this reason, the character's name comes second (*Figure 2. Title Focus: Miul the Hajduk. A Page from the Phanariot Era*: left).

The same thing happens for titles where the status appears associated with a placeEntity. Such a case can be noted for the titles *Santinela de la Grivița* [*The Sentinel from Grivița*] and *Moșneagul de la munte. Roman din Pind* [*The Old Mountain Man. Novel from Pind*]. The responses were similar across all age categories. The respondents do not consider space as relevant (*Pind*, *Grivița*, *mountain*), but turn their attention to status instead (*sentinel*, *old man*). An identical situation is noticed even for titles recording two characters associated with two different social statuses. For instance, in a title such as *Logodnica princepelui Matei Basarab* [*The Betrothed of Prince Matei Basarab*] all age groups consider the status (*betrothed*, *prince*) more important than the *personEntity* (*Matei Basarab*). Obviously, one cannot overlook the fact that the respondents may have been influenced by the (self-)marking of these *personEntities*: in fact, they are in the syntactic position of subject whose predicate, not expressed in the title, is the entire text itself (*Figure 2*: The fiancée of Prince Matei Basarab: right).

As noted, the data also indicate a preference among feuilleton novelists for combining more than two entities in their titles. In the 66 titles, the *personEntity* is the center to which the rest of the entities, namely *placeEntity* and *otherEntity* or

another *personEntity*, are linked. In general, it can be said that the *personEntity* only bears the company of two entities at most, in various combinations such as *placeEntity* + *placeEntity* + *personEntity*; *personEntity* + *personEntity* + *otherEntity*; *personEntity* + *otherEntity* + *otherEntity*. As *Figure 3* shows, triplets are rarer, while the doublets *placeEntity* + *personEntity* (3 occurrences), *personEntity* + *personEntity* (4 occurrences) or *personEntity* + *otherEntity* (5 occurrences) indicate a discrete strategy of highlighting the centrality of the person/character even when not explicitly mentioned by including proper names in the title string. Similar observations were made by Patras et al (2021, 183), who noted the tendency for persons to appear in the vicinity of other entities (12% of titles) in all titles of the European novels included in the study, rather than together with places (4.6%) or another person (4.4%).

placeEntity: When addressing the issue of spatial identifiers in feuilleton titles, we need to consider what expectations and knowledge a nineteenth-century reader had about where places were or were not marked on the map. In the 66 annotated titles, I have identified a total of 14 entities, which can be classified into: a. *indeterminate spatial entities:* 9 spaces > "South America", "Dacia", "Portugal", "Orient", "Pind", "mountain" (two occurrences), "on the trail", "a throne" and *b. determinate spatial entities/known to the reader:* 5 spaces > "Constanța", "Pitești", "Grivița", "Rucăr", "Valachia". The two categories defined above are designed to set in motion different imaginative mechanisms, as the feuilleton consumer tends to also choose a novel according to the space that is closest to him or her or the space that is furthest away, spatially or temporally. So far, we see that indeterminate spatial entities are mostly "exotic", while determinate spatial entities, familiar to the reader, are mostly found in Wallachia.

Direct experience of the spaces—"Constanța" is better known than "Dacia" or "South America"!—was also confirmed in the answers given by the questionnaire respondents. For example, a first analysis of the graph (*Figure 4*) shows that in the title *Diamantul pierdut. Scene din viața coloniștilor din America de Sud* [*The Lost Diamond. Scenes from the Lives of South American Settlers*], the space (*South America*) does not hold interest. All age groups consistently chose the category *otherEntities* (*diamond*) and *personEntity* (*settlers*). This preference indicates the identifying function that the titles fulfil, thus making explicit from the outset the main

subject of the book (*Lost Diamond + life of the settlers*). As an entirely thematic, literal title (Genette 1997, 81–82), it refers not only to the central theme of the work or to the characters involved, but also to a possible search scenario, thus to a detective narrative. Of the 184 people involved in my experiment, only eight chose the space as representative and descriptive of the text of the novel (*The Lost Diamond of South America*). As Genette points out, the functions of paratext do not manifest themselves homogeneously in audience reactions or across historical periods. For example, one person in the age group 15–20 selected, instead of a single keyword, three distinct words that must have been perceived as equally important: *diamond*, *settlers*, *South America*. The same happens with titles in which a *personEntity* appears in relation to a *placeEntity*. For example, for a title such as *Moșneagul de la munte. Roman din Pind* [*The Old Mountain Man. Novel from Pind*], the character is strongly marked, in contrast to the indeterminacy of spaces, either because of the vagueness of the location (*mountain*) or because of the Romanian readers' lack of familiarity (*Pind*).

Similar *phenomena* were observed when I analysed the occurrences of defined spaces, i.e. spaces that are familiar or, presumably, known to the readers. For a title such as *Radu de la Pitești* [*Radu of Pitești*], the entity designating the space (*placeEntity*=Pitești) was selected by the respondents in equal measure with the *PersonEntity=Radu*. The same is found in titles such as *Santinela de la Grivița* [*The Sentinel from Grivița*], *Drama din Constanța sau Taina unei mame. Roman realist, petrecut în zilele noastre [Drama from Constanța or A Mother's mystery. Realistic Novel Set in the Present Day*], *Din Rucăr. Scene văzute* [*From Rucăr. Scenes Witnessed*], *Olteanca cea frumoasă sau Inchiziția în Valachia în secolul al XIII lea. Roman istorico dramatic inedit, de mare actualitate* [*The Beautiful Woman from Olt or The Inquisition in Wallachia in the 13th century. A Novel of Historical Drama, Highly Up-to-Date*].

By considering the entities' map-ability, I defined localizable (*Portugal*, *Pind*, *South America*, *Constanța*, *Pitești*, *Grivița*, *Rucăr*) or non-localizable (*Dacia*, *Orient*, *mountain*, *on the trail*, *a throne*) *placeEntities*. There is a spread of indeterminate, exotic and imaginary spaces. Mentioning them in the titles of the feiulleton is, of course, a strategy to stimulate the imagination of the nineteenth-century reader who, except for the foreign news that might have mentioned them, had little reference to South America, Portugal or Pind. At the other end of the spectrum, the determined spatial entities show that the narratives are linked to the area of Wallachia (*Figure*

5), therefore we might presume that, alongside with the development of the central press in the capital of the young Kingdom of Romania (through the union of the Principalities of Moldavia and Wallachia), the evolution and development of the feuilleton novel is "a Wallachian business". If we analyze the list of approximately 650 titles published between 1850 and 1920 (extracted from the DCRR), we will rather notice that, among the approximately 368 novels first published in periodicals, 33 contain specific spatial entities that project the reader into the Wallachian space. [9]

As can be seen in *Figure 5*, the determined spatial entities are frequently located in the area of Muntenia/Wallachia (one occurrence), in the capital (nine occurrences: *Spirii Hill*, *Popa Chițu Street*, *Zalomit Street*) or in its approximate vicinity (*Mogoșoaia*, *Cozia*, *Grivița*, *Afumați*, *Pitești*, *Rucăr*, *Constanța*, *Valea Albă*). It is interesting to note that the occurrences of the spatial entities associated with Wallachia (and its capital, Bucharest) are recorded only in titles published after 1861 (Radu Ionescu, *Don Juanii din București* [*The Don Juans of Bucharest*]), that is, immediately after the Union of the Romanian Principalities in 1859. The need to clearly mark The Principality of Wallachia as a place where novel and surprising events are happening—we already know about the relocation of the capital to Bucharest, along with all the publications, institutions and personalities, once important in Moldavia—may be a symptom of the fact that unification idea (consequently, the national idea) was deeply imprinted in the collective imagination that nurtured the feuilleton novel. After 1866, when the Romanian Principalities are officially called Romania under the reign of Carol I, the titles of the feuilletons also start to reference this change of status (for instance, Constanța Dunca-Schiau, *O familie din București. Iezuiții României* [*A Family from Bucharest. The Jesuits of Romania*], 1867). A simple count shows that the entity *Romania* occurs eight times in all titles, immediately after *Bucharest.*

Titles also include references to Moldova (*Schitul Frumoasa* and *Moldova*) and Transylvania (*Pârlești*), but they indicate that this area had become a historical artefact, thus has lost its actuality relevance. The three occurrences of Moldavia are placed around character statuses (*Cantemir* and *Alexandru Lăpușneanu*, both of them medieval rulers of Moldavia), or to the historical event (*the fall of Moldavia*). It is noteworthy to add that these titles of nineteenth-century fiction already deliver an

[9] The 5 out of 66 titles reviewed included a total of 33 novels. A full list of these can be reviewed in the appendices.

idyllic image of Moldavia, that will be developed into a site of nostalgic memory in the writings of twentieth-century Moldavian writers of "Viața Românească" Circle. Below, I provide a few examples: *Alexandru Lăpușneanu, domnul Moldovei (1553–1568). Roman istoric* [*Alexandru Lăpușneanu, Ruler of Moldavia (1553–1568). Historical Novel*], *Epoca lui Dimitrie Cantemir, domnul Moldovei (1674–1711)* [*The Age of Dimitrie Cantemir, Ruler of Moldavia (1674-1711)*], *Cu paloșul. Poveste vitejească din vremea descălecatului Moldovei* [*With the Broadsword. A Brave Tale from the Time of the Conquest of Moldavia*].

On the contrary, the parodic title of Ion Slavici—*Revoluția din Pârlești. Roman tragodic în cinci părți, scos din niște manuscripte de la Ispania și alcătuit după cel mai nou sistem, dat fiind întru cinstea și mărirea celor întru toate deopotrivă între sine*—includes a somewhat determined space that restores the memory of a historical episode happened in Transylvania. The village *Pârlești*, mentioned in the title, is not a space that can be located on the map, but a fictitious name for the events of the spring of 1873 in Păuliș. *Revoluția de la Pârlești* [*The Pârlești Revolution*] brings to the public's attention the satirized image of the conflict between the local Hungarian and Romanian authorities.

Title structure: in the analysis of title structure, I focused on three aspects—the frequency of titles and subtitles, the connectors between titles and subtitles, and the presence/absence of the generic label in the (sub)title's content. The results show that more than half of the feuilleton titles also have subtitles, *i.e.* 47 out of 66. This confirms that titling practices do not resort to ambiguity tricks, but they rather mark a tendency towards explicitness and descriptiveness. Nineteenth-century novelists seem to prefer thematic titles because they announce the subject of the book and give the impression of a narrative continuity between the subject (the title) and the predicate (the text) of the oeuvre. The heterogeneity of the titling word strings also corresponds to subtitling practices of explication; accordingly, subtitles which may contain clues about space, events, characters, etc. Subtitles are linked to titles by distinct procedures such as: a. *interpunctuation* (42/ 66), b. *word-units* (3/ 66) or c. mixed: word-units and interpunctuation (2/ 66). The last case is often found in titles with a large number of entities, whereas interpunctuation is common to the majority of titles in the compiled list.

More than half of the novels (44/66) contain a (sub)genere label, which indicates the richness of novel sub-genres delivered in the feuilleton format. Being a paratextual element characterized by quasi-autonomy, the (sub)genre mark is included either in the title or in the subtitle; however, I also noted that there are titles consisting of a single word that do not associate a subtitle: e.g. N.G. Rădulescu-Niger: *Măriuca*, *Foamete* [*Famine*], *Fuga* [*The Escape*] or Anonymous: *Orfanii* [*The Orphans*]. I should also point out that the previously-provided examples remained exclusively in feuilleton-format, that is, they were not subsequently reprinted in volume; considering the explicitness of the surrounding headlines, fact that these titles lack a subtitle—therefore, they do not provide the readers with an explanation of the subject in the same fashion that newspaper headlines do—looks like an experiment in titling.

Among the 44 feuilletons containing a genre label, there is a rich variety of references such as *sensational novel*, *historical novel*, *dramatic novel*; additionally, they might also contain indications regarding the narrative form such as *novel*, *novel fragment*, *novella*, *story*, *tale*, *account*, *scene*, *episode*. During the nineteenth-century, the novel was still a genre not entirely defined and easily suspectable of faulty "originality"— because among the so-called "originals", sneaked a lot of localized translations (Teliban 2024) —, one can notice the hesitation in the titling practices regarding the use of genre labels, chiefly between *novel* (12 entries), *original novel* (6 entries) and *original novella* (5 entries). The data from my research aligns with the results obtained from the ELTeC titles. Apparently, the European literary space is divided into three title traditions: Germany and most Latin countries (France, Italy, Romania) favor the genre label *roman* (derived from *romance*), the Slavic-speaking territories prefer the label *povest* (highlining the narrative character), while England and Spain favor the term *novel(a)* (emphasizing the subject's novelty) (Patras *et al*. 2021, 177).

Turning to the contemporary reader's responses, we must inquire whether he/she recognizes and is sensitive to the sometimes highly creative formula/ phrasing of the (sub)genre naming. In most cases, for all age groups, there is hesitation about the choice of subgenre label. For titles such as *Mircea the Elder. Original Historical Novel*, the respondents oscillated between *novella*, *historical novel* and *original novel*, the latter two being chosen the most. The same is true for titles such as *Ursita. Nuvelă istorică* [*Ursita. Historical Novella*], *Lupta pentru mărire. Mare roman*

dramatic inedit [*Struggle for Greatness. Great Unpublished Dramatic Novel*] or *Omul muntelui. Roman original* [*The Mountain Man. Original Novel*], cases in which the respondents recognized the genre label by association with master genres (*dramatic novel*) or various novel attributes (*unpublished novel*, *original novel*, *historical novel*) rather than by taking them individually (*novel*, *novella*). This observation confirms the authors' tendency to make the title explicit, even in instances of creative articulation of the subgenre mark. The twenty-first century reader, in his/ her turn, reacts to subgenre explicitness, chiefly when originality and novelty are also warranted: *new novel*, *original (national) novel*, *unique novel*.

Developments in the titling practices of the Romanian feuilleton

Because the structure of the POP-LITE corpus allows for both subgenre and author inquiry, I decided to track the status and evolution over time of the entities in the titles of the most represented authors: 3 titles by D. Caselli, 7 titles by N.D. Popescu, 7 titles by N.G. Rădulescu-Niger. The titles of Caselli's novels are characterized by a decrease in the number of entities over time from 5 entities in 1896 (*placeEntity* + *personEntity* + *otherEntity* + *otherEntity* + *otherEntity*), to 2 entities in 1899 (*personEntity* + *otherEntity*). Behind this scheme, we may suspect a variation in all authors' use and distribution of entities: a title such as *The Beautiful Woman from Olt or Inquisition in Wallachia in the 13th century. A Highly Actual, Unpublished Historical-Dramatic novel* (1896) includes a complex, and rather ostentatious attribute combination, while *Ileana. A Novel of Great Sensation* (1899) simplifies the scheme. A distinct phenomenon can be observed in the novels authored by N.D. Popescu. The mention of entities (of all types, i.e. *personEntity*, *placeEntity*, *otherEntity*) shows some variation in number and time. Even if the data is extremely scarce, the great literary entrepreneur prefers to title his works unequivocally, especially his hajduk pieces, already forming a critical mass for subgenre emulation by the end of the nineteenth century. From 1864 to 1889, N.D. Popescu generally composes titles by observing simple recipes: two entities for the historical novels of 1864 and 1875; between three and four entities for the hajduk novels of 1881 and 1882; between two and three entities for the historical novels of 1887 and 1889. In the case of N.G. Rădulescu-Niger, I remarked a relative balance in the number of entities mentioned in the titles and variation between their type and association: *otherEntity+otherEntity* (1890), *personEntity+personEntity* (1891), *personEntity*

(1888, 1893 and 1896), *otherEntity* (1896). One can easily perceive the novelist's preference for the use of a single entity, thus a refusal of explicitness, as if, beyond the ephemeral condition of the feuilleton, the feuilletonist imagined a canonical destiny for his works.

The decrease in the number of entities is not a feature specific only to the Romanian feuilleton novel, as the shortening of titles seems to be a general trend, manifested throughout Europe. Studies such as those signed by Franco Moretti have suggested that titles have become shorter over time either for pragmatic reasons (imposed by editors and publishers), or out of a desire to convey the subject of the book to the public more succinctly (Moretti 2009, 134–158). While in the Middle Ages or the sixteenth, seventeenth and even sometimes eighteenth centuries, authors resorted to "synopsis" titles, which consisted of a summary of the action of the novel (e.g. *The Life and Strange Surprizing Adventures of Robinson Crusoe, of York, Mariner: Who Lived Eight and Twenty Years, All Alone in an Un-inhabited Island on the Coast of America, Near the Mouth of the Great River of Oroonoque; Having Been Cast on Shore by Shipwreck, Wherein All the Men Perished but Himself*), in the early nineteenth century these were replaced by short titles (Genette 1997, 71–72). Yet how does the trend towards making the title of the feuilleton explicit (described above) reconcile with a more general tendency to shorten titles, to reduce them to the bare minimum of words? Is this trend a copycat practice of newspaper headline writing practices?

Of course, the phenomenon can be explained by the relationship between the feuilleton novel and the journalistic genres of the nineteenth century. Between these pieces of text conveyed in different functional styles there is no direct communication at the level of content; but there is a subtle communication generated and supported by the newspaper template through the use of miscellaneous headings (*Proverbs and Sayings*, *Questions and Answers*) that highlight the action or situation in which the characters of the novel find themselves. The idea of making the feuilleton title explicit is also specific to journalistic genres such as the premier-Paris (front-page article), the chronicle, the interview, the report (Thérenty 2007, 207). However, in newspaper headlines, there is a focus on *otherEntity* rather than *personEntity* so, compared to feuilleton fiction, the distribution of entities is inverse. For example, the novel *Fiica lui Hagi Dumitru* [*Hagi Dumitru's Daughter*] by Rădulescu-Niger contains two co-dependent *personEntities*. In the issue dated

December the 8th 1891, of the newspaper "Romania. Organ Liberal-Conservator", the novel was published alongside polemical articles (*Guvern de contrabandă* [*Smuggling Government*]), short or investigative reports (*Conflictul franco-bulgar* [*The Franco-Bulgarian Conflict*], *Criza din Austria* [*The Crisis in Austria*]) and even a column entitled *Dădacile*, which mocks the influence of the nannies on the decisions of political leaders. A quick glance at newspaper headlines shows that they mainly contain *otherEntities* (*conflict*, *government*, *crisis*) and a *placeEntity* (*Austria*). In the newspaper "Adevărul Politic" issued in March the 16th 1904, the novel *The Old Mountain Man. Novel from Pind* by Brun Jules (N. Papahagi)—centred on *personEntity*—is surrounded by other headlines that often record otherEntities (*Întrunirea partidului conservator* [*The Conservative Party Meeting*], *Întrunirea conservatorilor* [*The Conservative Meeting*], *Războiului ruso-japonez* [*Russian-Japanese War*]) or combinations of *placeEntities* and *otherEntities* that might sound as feuilleton titles (*Crima din str. Atelierului* [*Murder on Atelierului Street*], *Sinuciderea de la hotel Moscova* [*Suicide at the Moscow Hotel*]). However, there are also cases in which speeches by members of parliament (*Discursul lui Take Ionescu* [*Take Ionescu's speech*]) or articles of political polemic (*Bir pe țărani [Tax on the Peasants]*) are reprinted. This time, the *personEntities* (*Take Ionescu*, *the peasants*) are not informative as they depend on the *otherEntity* (*the speech*, *tax*).

Conclusions

Composed of a maximum of two entities, the titles of the feuilleton novels use as a frequent formula the combination *personEntity + (any) other entity* (*placeEntity + personEntity, personEntity + personEntity, personEntity + otherEntity*), present in 33% of the 66 annotated titles. The combination found here confirms the broader research applied to ELTeC titles: throughout the nineteenth century, novel titles rely on the mention of at least two entities, one of which is necessarily a person (Patras et al 2021, 186). As far the use of the *personEntity* is concerned, the comparison of the feuilleton titles and journalistic headlines indicates: a. the importance of the character as a cohesive element of a sequenced narrative; b. the embedding, in the feuilleton code, of an engagement reading. Considered less cohesive for the narrative, spaces rarely appear in the titles of the feuilleton; however, their mentions, whether definite or indefinite for the readers, echo recent historical events such as the union of the Principalities, the proclamation of the Kingdom, the relocation of

social, economic and cultural life to Bucharest. In short, there is a media matrix shared by both press and feuilleton fiction, which means that the topical and miscellaneous facts sought by an eye formed through the newspapers' daily reading will also be tracked in the feuilleton titles (Thérenty 2007, 113).

In 71% of all titles, I noticed a trend to make the title more explicit by adding a subtitle, whether represented by a genre label or not. However, the case studies carried out on the feuilletons authored by N.D. Popescu, D. Caselli, N.G. Rădulescu-Niger have shown that each novelist has his or her own recipe for the title, influenced to a greater or lesser extent by the pressure of the typographical format or the tendencies of the newspaper template.

As we have only a few testimonies about the reading "with bated breath" of consumer novels, published in feuilleton format (Sadoveanu 1957, 550; Pillat 1978, 63-108; Pillat 1998, 128–144), I thought that a *reader-survey* experiment could reveal the stimuli (keywords) to which readers often react. I discovered that the audience finds status more stimulating than the character's name—for instance, in the phrase "Prince Dimitrie", *Prince* is chosen more often than *Dimitrie*—which opens up a horizon for sociological reflection (the feuilleton novel represents and addresses a world marked by social differences). A social status means, in fact, a concentrated narrative, the story of a person's arrival to a certain social condition, thus it is considered more defining and more stimulating than a bare name.

The results obtained are only a starting point for researching the relationship between the relative precision of the title and the relative improvisation of the feuilletonistic writing, between the title as theme and the episodes of the feuilleton as variations. Together with these aspects, the present study aimed to apply and adapt a previous annotation model for titles, as well as to introduce an analysis methodology that calibrates the excesses of distant reading by adding an empirical, cross-checking approach inspired by recent reader-response surveys.

Annex 1: Figures

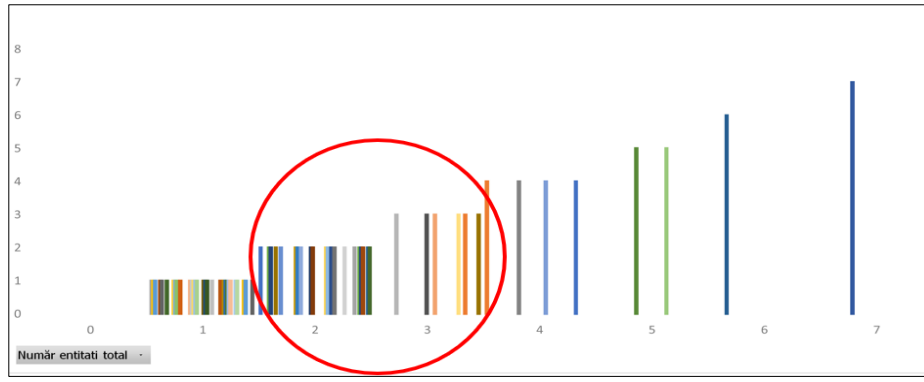


Figure 1: Overall distribution of entities

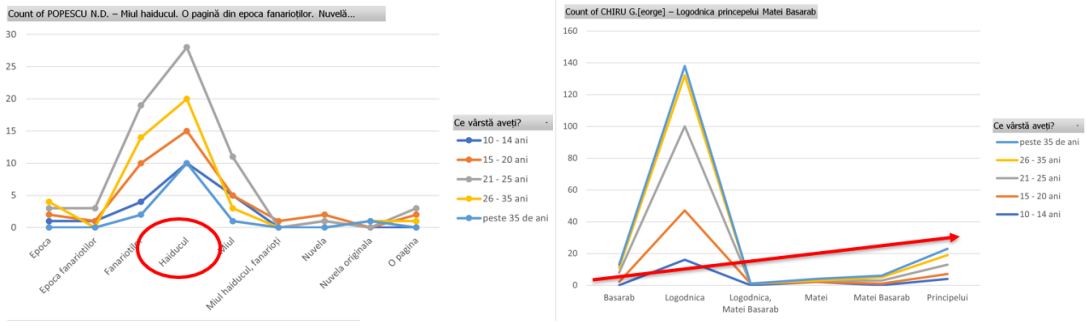


Figure 2: Title Focus: Miul the hajduk. A page from the Phanariot era (left)
The fiancée of Prince Matei Basarab (right)

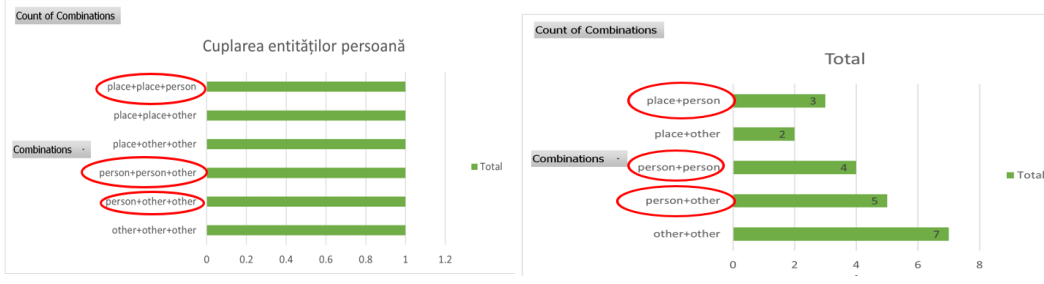


Figure 3: Pairing of PersonEntities in threes vs. twos

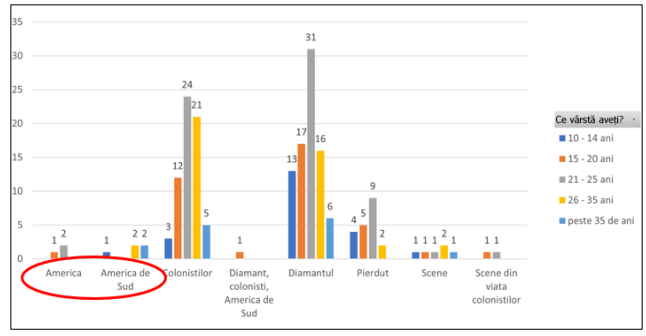


Figure 4: Title Focus: The Lost Diamond. Scenes from the Lives of South American Settlers

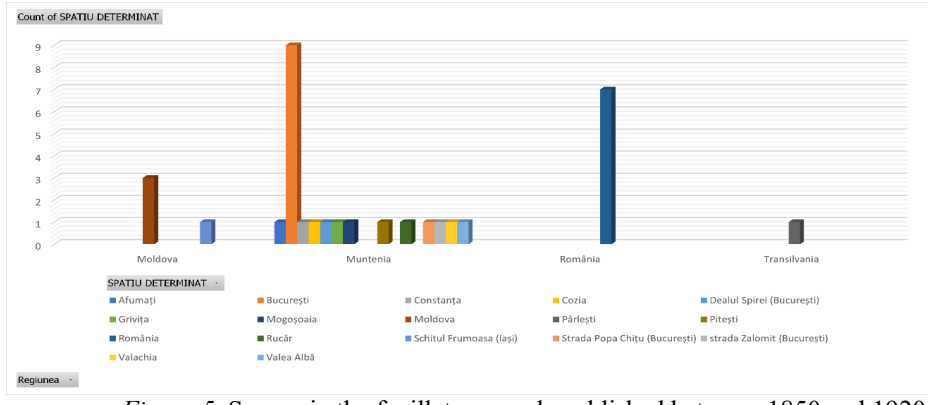


Figure 5: Spaces in the feuilleton novels published between 1850 and 1920

Annex 2: The list of novels contains restored titles, with the date of first feuilleton publication

AUTOR	TITLU
LUNGU Emilia	*Elmira. Roman original în trei părți* (1882)
ALLAN George [Mite KREMNITZ]	*Prințul Dimitrie. Roman* (1883)
RĂDULESCU-NIGER N. G.	*Căpitanul Ropotă* (1886)
RĂDULESCU-NIGER N. G.	*Din memoriile unui procuror* (1888)
POPESCU N.D.	*Santinela de la Grivița* (1887)
RĂDULESCU-NIGER N. G.	*Foamete* (1896)
RĂDULESCU-NIGER N. G.	*Fuga* (1896)
HASDEU Iulia	*Domnișoara Ursuza. Roman inedit pentru copii scris la 1881* (1891)
POPESCU N.D.	*Miul haiducul. O pagină din epoca fanarioților. Nuvelă originală* (1871)
POPESCU N.D.	*Tunsul haiducul. Nuvelă originală ilustrată cu patru gravuri* (1870)
POPESCU N.D.	*Bujor haiducul. Nuvelă originală ilustrată cu patru gravuri* (1872)
FULGER Const. [Vasile POP]	*Revoluția din Portugalia sau Un tron pentru iubire. Mare roman de actualitate senzațională* (1910)
POPESCU N.D.	*Radu al III-lea cel Frumos. Novelă originală* (1864)
POPESCU N.D.	*Mircea cel Bătrân. Nuvelă istorică originală* (1876)
POPESCU N.D.	*Sora de caritate. Un episod din războiul de independență* (1880)
MACRI Panait	*Studentul. Roman original* (1884)
RĂDULESCU-NIGER N. G.	*Fata lui Hagi Dumitru* (1891)

AUTOR	TITLU
RĂDULESCU-NIGER N. G.	*Măriuca* (1896)
RĂDULESCU-NIGER N. G.	*Gelozie. Roman* (1890)
XENOPOL Adela	*Dorin Molna* (1893)
ESCHENAZI Maria	*Estella și Maria sau Răsplata virtutei* (1894)
LECCA Irina G.	*Pe urma dragostei. Roman* (1909)
CASSVAN Sarina	*Între artă și iubire. Roman* (1918)
MĂRUNȚEANU Grigore	*Contrast. Fragment dintr-un roman, de sub presă* (1895)
VULCAN Petru	*Dragomir* (1895)
LUPU E. [Zamfir C. ARBORE]	*Diamantul pierdut. Scene din viața coloniștilor din America de Sud* (1893)
MOLDOVAN Vasile E.	*Dușmănie. Roman* (1903)
BRUN Jules, N. PAPAHAGI	*Moșneagul de la munte. Roman din Pind* (1904)
HASDEU Bogdan Petriceicu	*Ursita. Nuvelă istorică* (1864)
ANDRIESCU BOGDAN N.	*Trandafir. Roman original* (1878)
POPESCU Iosif	*Petru Cazacul* (1879)
RUSSU Ioan	*Nicu Ciobănelul. Nuvelă din popor* (1887)
ZORILĂ Nic. T.	*Vezinas. Istorisire din timpul lui Duras* (1892)
CLOROFORM [Grigore MĂRUNȚEANU]	*Sfinx sau din secretele unei nopți tainice. Roman misterios în trei capitole întunecate* (1899)
Aslan Edgar	*Fiica lui Radu cel frumos. Împrejurări și chipuri din trecut* (1905)
OECONOMU Ciru	*Fiica lui Sejan. Poveste istorică* (1898)
CHITUL Tit	*Catastrofa Daciei. Roman istoric* (1903)
CHITUL Tit	*Steaua Orientului. Roman istoric* (1905)

AUTOR	TITLU
DE LA BISTRIȚĂ	*Voievozii. Povestire originară a „Unirii”* (1912)
anonim	*O resbunare nouă* (1865)
ARGHIROPOL Emanoil	*Condamnata. Tăcere in viaţă, iertare în ora morţii* (1868)
FUNDESCU I.C.	*Radu de la Pitești* (1891)
CASELLI D.[omenico]	*Lupta pentru mărire. Mare roman dramatic inedit* (1896)
CASELLI D.[omenico]	*Ileana. Roman de mare senzație* (1899)
anonim	*Copilașii Marioarei. Roman senzațional* (1905)
NICOLESCU WOLLMAR D.	*Triumful iubirii* (1910)
DOAMNA L. [Marie BOUCHER, V.A. URECHIA]	*Omul muntelui. Roman original* (1857)
FATIMA	*Zoe sau După divorț* (1892)
DRÂNCEANU Ion	*Princesa* (1893)
OECONOMU Ciru	*Din Rucăr. Scene văzute* (1903)
VAIAN Eugen	*Măritată! Roman nou* (1893)
MIHĂESCU-NIGRIM N.	*Amor ascuns. Roman original* (1899)
VULCAN Petru	*Almeiuna (Din Orient). Roman* (1903)
PETRAPETRESCU Horia	*Barcarolă. Un capitol dintr-un roman original* (1907)
NUTZESCU Const.	*Un vis de-o clipă. Roman* (1910)
ANDREI I.	*Șeina. Roman original* (1911)
MESTUGEAN V.	*Getta. Roman senzațional* (1916)
DIMANCEA G.I.	*Alexandru. Roman original* (1882)
NEGRU Mihail	*O poveste de iubire* (1911)
POP Vasile	*„Noi suntem sclave”. Roman modern* (1912)

AUTOR	TITLU
anonim	*Orfani!* (1897)
anonim	*Pentru moștenire* (1897)
anonim	*Drama din Constanța sau Taina unei mame. Roman realist, petrecut în zilele noastre* (1916)
anonim	*Femeie fără suflet. Mare roman dramatic* (1916)
CHIRU G.[eorge]	*Logodnica princepelui Matei Basarab* (1894)
CASELLI D.[omenico]	*Olteanca cea frumoasă sau Inchiziția în Valachia în secolul al XIIIlea. Roman istoricodramatic inedit, de mare actualitate* (1896)

Annex 3: List the 33 out of 368 novels of the feuilleton novels containing definite spaces

AUTOR	TITLU	AN	SPATIU DETERMINAT	REGIUNEA
IONESCU Radu	*Don Juanii din Bucureşti*	1861	București	Muntenia
DUNCA-SCHIAU Constanța	*O familie din Bucureşti. Iezuiții României*	1867	România	România
			București	Muntenia
DUNCA-SCHIAU Constanța	*Sub vălul Bucureştilor. Fiica adoptată*	1868	București	Muntenia
MISSAIL George	*România veche și România nouă. Roman original de datini și moravuri naționale*	1869	România	România
M.	*O săptămână la schitul Frumoasa*	1871	Schitul Frumoasa (Iași)	Moldova
SLAVICI Ioan	*Revoluția din Pârlești. Roman tragodic în cinci părți, scos din niște manuscripte de la Ispania și alcătuit după cel mai nou sistem, dat fiind întru cinstea și mărirea celor întru toate deopotrivă între sine*	1873	Pârlești	Transilvania

AUTOR	TITLU	AN	SPATIU DETERMINAT	REGIUNEA
POPESCU N.D	*Don Juanii Bucureştilor. Jujuca*	1875	București	Muntenia
POPESCU N.D.	*Alexandru Lăpuşneanu, domnul Moldovei (1553–1568). Roman istoric*	1875	Moldova	Moldova
POPESCU N.D.	*Fata de la Cozia. Nuvelă istorică originală, precedată de un rezumat al domniei lui Vlad-Vodă Dracula*	1877	Cozia	Muntenia
XENOPOL Nicolae D.	*Păsurile unui american în România*	1879	România	România
GRANDEA Grigorie H.	*Berlicoco. Istorie scrisă cu ajutorul dracului din România*	1880	România	România
POPESCU N.D	*Botezul de sânge. Nuvelă istorică originală, Episoade din istoria resbelului independenții României*	1882	România	România
POPESCU N.D.	*Radu al VII-lea de la Afumați (1521–1529). Nuvelă istorică*	1885	Afumați	Muntenia
CREȚEANU George (?)	*Misterele din Bucureşti*	1886	București	Muntenia

AUTOR	TITLU	AN	SPATIU DETERMINAT	REGIUNEA
POPESCU N.D.	*Epoca lui Dimitrie Cantemir, domnul Moldovei (1674–1711)*	1886	Moldova	Moldova
POPESCU N.D.	*Santinela de la Griviţa*	1887	Griviţa	Muntenia
FUNDESCU I.C. (?)	*Dramele Bucureştilor*	1888	Bucureşti	Muntenia
FUNDESCU I.C.	*Scheletul din Strada Popa Chiţu nr. 17*	1889	Strada Popa Chiţu (Bucureşti)	Muntenia
MACRI Panait	*Cea mai frumoasă fată a Bucureştiului. Roman original de mare senzaţiune*	1889	Bucureşti	Muntenia
FUNDESCU I.C.	*Radu de la Piteşti*	1891	Piteşti	Muntenia
CASELLI D.[omenico]	*Olteanca cea frumoasă sau Inchiziţia în Valachia în secolul al XIIIlea. Roman istoricodramatic inedit, de mare actualitate*	1896	Valachia	Muntenia
M.B.	*Pruncucigaşa. Drama din strada Zalomit. Roman din zilele noastre*	1896	strada Zalomit (Bucureşti)	Muntenia
CASELLI Dom.[enico]	*Din începutul Bucureştilor*	1898	Bucureşti	Muntenia

AUTOR	TITLU	AN	SPATIU DETERMINAT	REGIUNEA
OECONOMU Ciru	*Din Rucăr. Scene văzute*	1903	Rucăr	Muntenia
BECESCUSILVAN Gh.	*Valea Albă, 1476. Roman războinic*	1904	Valea Albă	Muntenia
Rosetti Radu	*Cu paloşul. Poveste vitejească din vremea descălecatului Moldovei*	1905	Moldova	Moldova
anonim	*Mamzella din Dealul Spirei (fragment)*	1909	Dealul Spirei (București)	Muntenia
FRUNZĂ G. [Constantin MILLE]	*Crimele Bucureştilor. Mare roman popular de senzație*	1909	București	Muntenia
FRUNZĂ G. [Constantin MILLE]	*Drama din staţia Mogoşoaia. Crimă ori sinucidere? Mare roman de senzație*	1909	Mogoșoaia	Muntenia
FULGER C. [Vasile POP]	*Grozăviile poliției noastre sau România în fața unui război din cauza unor documente furate. Roman senzațional*	1909	România	România
FULGER Const. [Vasile POP]	*Dramele staţiunilor balneare din România. Mare roman de*	1911	România	România

AUTOR	TITLU	AN	SPATIU DETERMINAT	REGIUNEA
	actualitate senzațională,			
anonim	*Drama din Constanța sau Taina unei mame. Roman realist petrecut în zilele noastre*	1915	Constanța	Muntenia

References

Anghelache, Constantin, Iordan Petrescu, Mădălina Gabriela Anghel, and Emilia Gogu. 2018. *Evoluția centenară a învățământului în România*. București: Editura Economică.

Cavallo, Guglielmo, and Roger Chartier (*eds*). 2003. "New Readers in the Nineteenth Century: Women, Children, Workers." In *A History of Reading in the West*, 313–344. Amherst: University of Massachusetts Press.

Dicționarul cronologic al romanului românesc de la origini până în 2000 (DCRR). 2023. Vol. I–II. Ediție revizuită și adăugită. Cluj-Napoca: Presa Universitară Clujeană.

Genette, Gérard. 1997. *Paratexts: Thresholds of Interpretation*. Translated by Jane E. Lewin. Cambridge: Cambridge University Press.

Lyons, Martyn. 2003. "New Readers in the Nineteenth Century: Women, Children, Workers." In *A History of Reading in the West*, edited by Guglielmo Cavallo and Roger Chartier, 313–344. Amherst: University of Massachusetts Press.

Moretti, Franco. 2009. "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740–1850)." *Critical Inquiry* 36 (1): 134–158.

Patraș, Roxana, Carolin Odebrecht, Ioana Galleron, Rosario Arias, Berenike J. Herrmann, Cvetana Krstev, Katja Mihurko Poniž, and Dmytro Yesypenko. 2021. "Thresholds to the 'Great Unread': Titling Practices in Eleven ELTeC Collections." *Interférences Littéraires/Literaire Interferenties* 25: 163–187. URL: http://interferenceslitteraires.be/index.php/illi/article/view/1102. Accessed May 2021.

Patraș, Roxana. 2018. *The Remains of the Day: Literature and Political Eloquence in 19th-century Romania*. Rome: Aracne editrice.

Patras, Roxana. 2022a. "WORKFLOW - STAGE 1 of the POPLITE Project." URL: https://doi.org/10.5281/zenodo.7360200. Accessed February 28, 2023.

Patras, Roxana. 2022b. "WHITE PAPER of the POPLITE Project." URL: https://doi.org/10.5281/zenodo.7360195. Accessed February 28, 2023.

Pillat, Dinu. 1978. "Romanul de senzație în literatura română din a doua jumătate a secolului al XIX-lea." In *Itinerarii istorico-literare*, 63–108. Bucharest: Minerva.

Pillat, Dinu. 1998. "Cum se deschide la noi gustul pentru romanul senzațional." In *Mozaic istorico-literar. Secolul XX*, 128–144. Bucharest: Albatros.

Proiectul Pop-Lite. URL: https://poplite.org/despre/. Accessed February 28, 2023.

Sadoveanu, Mihail. 1957. *Opere. Vol. IX: Demonul tinereții. Olanda. Împărăția apelor. O întâmplare ciudată*. Bucharest: Editura de Stat pentru Literatură și Artă.

Sampling criteria for the ELTeC. 2022. URL: https://distantreading.github.io/sampling_proposal.html. Accessed May 25, 2022.

Schmidt, Siegfried J. 1982. *Foundation for the Empirical Study of Literature: The Components of a Basic Theory*. Translated by R. de Beaugrande. Hamburg: Helmut Buske.

Slavici, Ioan. 1873. *Revoluția din Pârlești. Roman tragodic în cinci părți, scos din niște manuscripte de la Ispania și alcătuit după cel mai nou sistem, dat fiind întru cinstea și mărirea celor întru toate deopotrivă între sine*. Published in *Gura satului*, XII, no. 14–24.

Teliban, Andreea. 2024. "La curțile documentului. Două proze franțuzești în traducere românească: *Catastihul amorului* și *La gura sobei*." *Observator cultural* nr. 1202/April 4. URL: https://www.observatorcultural.ro/articol/la-curtile-documentu-lui-doua-proze-frantuzesti-in-traducere-romaneasca-catastihul-amorului-si-la-gura-sobei. Accessed April 6, 2024.

Thérenty, Marie-Ève. 2007. *La littérature au quotidien. Poétiques journalistiques au xixe siècle*. Paris: Éditions de Seuil.

Exercises in Literary Geography: where hajduks roam in the second half of the nineteenth century[1]

Alexandra OLTEANU

Geocriticism and spatial criticism

Geocriticism and spatial criticism are two distinct but interconnected approaches used within the fields of literary and cultural studies. Both approaches focus on the relationship between literature, space and place, but they differ in their theoretical foundations and emphases. Geocriticism, as a term, was first coined by French geographer Bertrand Westphal in the early 2000[s] (Westphal 2011, 14). Geocritics examine the ways in which literature engages with space, and they emphasize the geographical imagination and the representation of landscapes in literary texts. One of the foundational ideas of geocriticism is the notion of *topophilia*, introduced by geographer Yi-Fu Tuan; it refers to the love of place (Tuan 1990, 12). In Tuan's examination of topophilia, he elucidates the diverse methods through which individuals formulate a sense of sites and establish connections with their surroundings. This involves the cultivation of positive feelings and attitudes towards specific spaces. Tuan posits that these affective bonds play an essential role in shaping human perceptions of and engagement with the environment. Topophilia, in its full comprehensive scope, extends beyond the physical attributes of a location to encompass cultural, historical, and the symbolic connotations attached to it. Tuan underscores that these emotional affiliations contribute significantly to the multifaceted nature of human encounters with the environment. An appreciation of

[1] This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS –UEFISCDI, project number PN-III-P1-1.1-TE-2019-0127, within PNCDI III.

topophilia facilitates a profound exploration of the subjective and emotional dimensions inherent in human-landscape relationships, thereby offering insights into the intricate interplay between individuals and their surroundings:

> The existence of a kinesthetic relationship between certain physical forms and human feelings is implied in the verbs we use to describe them: for example, mountain peaks and man-made spires "soar," ocean waves as well as architectural domes "swell" [...] Moreover, architectural forms appear to influence our impression of size—of the way space expands and contracts to a degree that natural landforms rarely give. (Tuan 1990, 28)

The narrative of this work is infused with the concept of topophilia, which encapsulates the emotional connection between individuals and their surroundings. The exploration spans a diverse array of experiences, ranging from the profound allure of a cathedral to the vibrant hustle and colors of a market stall, the majestic presence of the Alps, the rhythmic embrace of waves on a beach, and the nuanced composition of planes within a landscape garden. Yi-Fu Tuan's discerning observations extend to a myriad of worldly phenomena, from the massive to the microscopic that captivate his affectionate scrutiny (Tuan 1990, 78). Nevertheless, the scope of his inquiry transcends topophilia alone. It encompasses a broader spectrum of perceptions, attitudes, and values, intertwined with intricate depictions of landscapes and townscapes from various time periods. The book delves into topics such as the distinctions between human sensory experiences and those of other species, the systems we use to classify the world around us, our views of the afterlife and the future, the ways in which landscape and scenery are defined, and the dynamic connection between symbols and realities. Themes related to contrasts between the primitive and the modern, between wilderness and paradise, and between rural and urban continuously emerge, submerge, and resurface, akin to playful seals in an archipelagic sea (Tuan 1990, 82).

Geocritics argue that literature not only reflects real geographical spaces but also constructs and influences our perceptions of these spaces. Prominent geocritics,

such as Edward Said in his work *Culture and Imperialism*, analyze the role of literature in shaping cultural perceptions of colonial spaces. Said's postcolonial geocriticism explores how literature contributes to the creation of geographical imaginaries that influence power dynamics and identity formation (Said 1994, 18). Dane Kennedy's article "Imperial History and Postcolonial Theory" plays a significant role in the ongoing discourse about the relationship between literary criticism and historical studies, specifically within the context of the interaction between "colonial discourse analysis" and imperial historiography. His balanced intervention thoughtfully analyzes the shortcomings found in the approaches of (post)Saidian literary critics regarding their engagement with historical evidence and historiographical methods. Additionally, Kennedy laments the widespread missed opportunities among imperial historians to fully embrace what is implied by the emergence of postcolonial cultural studies (Kennedy 1996, 352).

Other theorists like Franco Moretti focus on spatial patterns in literature, using mapping and distant reading techniques to uncover trends and relationships across different texts. Moretti's methodology enables him to discern overarching literary trends and patterns spanning temporal and spatial dimensions. Through this approach, Moretti endeavors to offer a macroscopic perspective on literary history, unveiling concealed structures that may elude traditional analytical methods. The ideological ramifications of Moretti's scholarly interventions manifest in the tension between macroscopic analysis and individual agency. Detractors contend that an exclusive focus on statistical trends runs the risk of disregarding the intricacies and complexities inherent in individual texts and authors. This threatens to homogenize literature, potentially neglecting the distinct cultural and historical contexts intrinsic to each work. Moreover, the reliance on quantitative data prompts questions about the potential risk of reducing literature to mere data points, thereby overlooking the subjective and qualitative facets of artistic expression. Comparing Moretti's macroscopic approach and Said's historiographical viewpoint reveals both converging and diverging features. In *Orientalism*, Said critiques Eurocentric perspectives within Western scholarship, advocating for a more nuanced understanding of cultural texts (Said 1994, 55). Said's emphasis on understanding

power dynamics, imperialism, and the construction of otherness introduces a layer of complexity to the discourse regarding literary spaces. Applying Said's critique of orientalism to Moretti's work should raise relevant questions about whether distant reading inadvertently perpetuates a form of occidentalism by reducing diverse literary traditions to universalized patterns. Scholars have engaged in debates surrounding whether Moretti's methodologies risk perpetuating the very hegemonic structures he aims to expose.

Hajduks and the space

As compared to geocriticism, spatial criticism—sometimes referred to as literary cartography—has roots in human geography and critical theory. The spatial turn in literary studies gained momentum in the late twentieth century, with theorists like Henri Lefebvre and Michel de Certeau emphasizing the social and cultural dimensions of space. Spatial critics argue that space is not merely a backdrop for literary events but actually a dynamic force that shapes and is shaped by human actions. Henri Lefebvre's *The Production of Space* is another work that examines the social construction of space and its impact on everyday life. Lefebvre introduces the idea of the *spatial triad* to understand the connection between perceived space, conceived space, and lived space (Lefebvre 1991, 22). Spatial critics draw on Lefebvre's concepts to argue how literature reflects and challenges societal spatial practices:

> From the social standpoint, space has a dual "nature" and (in any given society) a dual general "existence." On the one hand, one (i.e. each member of the society under consideration) relates oneself to space, situates oneself in space. One confronts both immediacy and an objectivity of one's own. One places oneself at the centre, designates oneself, measures oneself and uses oneself as a measure. (Lefebvre 1991, 175)

Lefebvre asserts that space transcends mere physicality, and he posits that it is elaborately molded by societal behaviors, ideologies, and power dynamics. He

introduces the conceptual distinctions of *conceived space* (representations of space), *perceived space* (experiential dimensions of space), and *lived space* (the quotidian, tangible reality). Lefebvre contends that generating space constitutes a social phenomenon that entails both tangible and symbolic. Furthermore, he critiques the oversimplified view of space as solely governed by economic or technical considerations. This theory has left an indelible mark on the fields of urban studies, geography, and cultural studies, challenging established paradigms surrounding space and furnishing a theoretical framework for scrutinizing the social construction of the environment. Michel de Certeau, in *The Practice of Everyday Life*, further contributes to spatial criticism by introducing the concepts of space strategies and space tactics (De Certeau 1984, 95). De Certeau underscores the inventive nature of routine activities, highlighting how individuals, through their ordinary practices, establish realms of agency within prevailing systems. Of particular significance is the chapter dedicated to the concept of *walking in the city*, wherein de Certeau explores how pedestrians, by charting their unique paths and narratives, actively defy the predetermined order imposed by urban planning:

> Unlike Rome, New York has never learned the art of growing old by playing on all its pasts. Its present invents itself, from hour to hour, in the act of throwing away its previous accomplishments and challenging the future. A city composed of paroxysmal places in monumental reliefs. The spectator can read in it a universe that is constantly exploding. In it are inscribed the architectural figures of the *coincidatio oppositorum* formerly drawn in miniatures and mystical textures. On this stage of concrete, steel and glass, cut out between two oceans (the Atlantic and the American) by a frigid body of water, the tallest letters in the world compose a gigantic rhetoric of excess in both expenditure and production. (De Certeau 1984, 102)

Lefebvre and de Certeau present mutually reinforcing viewpoints on the interaction between space and everyday life. Lefebvre builds a macro-level theory, delving into the social production of space, while de Certeau creates a micro-level exploration, examining how individuals engage with and challenge spatial frameworks in their daily routines. The influence of both works is significant,

extending to cultural theory, urban studies, and broader understandings of the experiential dimensions of space. Spatial critics apply these ideas to literature, examining how fictional characters navigate and resist spatial constraints. Franco Moretti's spatial criticism entails examining how literature portrays space and how these representations contribute to understandings of social and cultural phenomena. The *Atlas of the European Novel 1800–1900* explores the geographic settings of European novels to glean insights into the complex connections between literature and geography:

> The young hero of *The Captain's Daughter*, for instance, who dreams of going to Petersburg, is promptly dispatched in the opposite direction, to the eastern periphery of the Czarist empire. In *Waverley*, Charles Stewart never completes his march towards London: he lands in the North-West of Scotland, raises the Standard of Rebellion in the middle of the Highlands, crosses the Highland line, reaches Edinburgh, crosses the Anglo-Scottish border, reaches Derby—and then stops. He stops, in other words, exactly where Austen's England begins (Pemberley, the northernmost locality "seen" in her novels, is also in Derbyshire). And that Scott's world should end exactly where Austen's begins, and Austen's end where Scott's begins. (Moretti 1999, 33)

Franco Moretti's examination of the spatial trajectories of characters in *The Captain's Daughter* and *Waverley* provides a perceptive analytical framework for investigating the dynamics of literary geography in British novels. In both instances, intentional narrative shifts redirect the characters away from their original aspirations, guiding them towards the peripheries or boundaries of their respective empires. The juxtaposition of the protagonist's aspiration to reach Petersburg in *The Captain's Daughter* with his journey to the eastern periphery of the Czarist empire not only signifies a divergence in the narrative but also implies a broader commentary on the sociopolitical milieu of the period (Moretti 1999, 32). Moretti emphasizes the strategic deployment of geographical settings as a tool for examining and commenting on power dynamics, emphasizing the peripheries as crucial sites for narrative exploration. Likewise, in *Waverley*, the decision to interrupt Charles Stewart's march towards London reflects a conscious selection of geographical scope.

Moretti posits that this purposeful interruption corresponds to the geographical boundaries of Austen's England, particularly spotlighting Pemberley in Derbyshire as the most northerly setting in her novels (Moretti 1999, 33).

The observation that Scott's fictional world concludes precisely where Austen's commences, and vice versa, prompts contemplation of the symbolic significance embedded in these geographical intersections. This observation raises questions about the thematic implications of the chosen locations, indicating a nuanced connection between distinct literary realms. Moretti's analysis acts as a catalyst, prompting readers to investigate the function of literary geography not solely as a setting but also as a deliberate narrative device that shapes the nuances of the fictional worlds meticulously crafted by these authors. The essential inquiry that instigated this investigation was whether Moretti's concepts pertaining to the functions of peripheral spaces and the thematic implications of locations could be used in case studies within Romanian literature. The HAIRO corpus serves as the primary source of material for such an examination (Patras *et al.* 2020).

In the genre of Romanian Hajduk novels, there exists a profound correlation between character and its geographic space. The Hajduk's ability to traverse various landscapes is grounded in the exploration of an extensive territory characterized by diverse topographies Elements such as evading authority, executing remarkable escapes, and cultivating a personal connection with a nurturing natural environment are portrayed in both the textual narrative and the series of illustrations integrated into the novels.

The illustrations in *Figures 1* and *2* depict a profound correlation between the central character of the Hajduk novel and the spatial context in which the narrative unfolds. The geographic settings of the Hajduks' exploits are intentionally chosen, serving to facilitate both the mobility of the noble outlaws and the sensational nature of their actions. The "engravings"/ illustrations capture instances of remarkable escapes, wherein the courage of the Hajduk Codreanu is heightened by spatial considerations. The act of leaping from a lofty cliff and vaulting over a tall wall—executing this feat amid the intense scrutiny of adversaries—reflects integral elements of a of a narrative relevant to literary geography.

I analyze the Hajduk novel as a hybrid subgenre that combines elements of historical, social, and adventure novels. These literary features reveal profound shifts in historical and cultural perceptions of space. The Hajduk novel illustrates transformative alterations in how individuals across diverse historical eras and cultures have conceived the world imaginatively. Hajduks establish a natural, feudal society in opposition to official authority, choosing the protective refuge of nature, specifically the forest, as the territory for shielding themselves from adversaries who were pursuing them. The city and the palace emerge as danger zones, where conspiracies against the Hajduks and the oppressed populace they safeguard are devised by corrupt rulers. The intrigues within the Hajduk novel attest that spatiality transcends the fact of being a mere framework or backdrop for the Hajduk's rebellion; rather, the space assumes a central role in the narrative's evolution. A potential explanation for the absence of the "great city" in the Hajduk novel is provided by Dinu Pillat in a commentary on an article authored by Al. A. Philippide. Philippide establishes a correlation between the scarcity of the adventure subgenre in Romania and the absence of native metropolises capable of serving as settings for sequences of unforeseen and captivating events:

> Al. A. Philippide, the only Romanian writer who questioned why the adventure novel was almost non-existent in our literary landscape, attributes this phenomenon to the absence of the "great city," a true "inextricable labyrinth" that "offers at every step unforeseen things, unexpected encounters, strange coincidences, infinite possibilities of disappearance, hard-to-discover hiding places, and numerous and varied characters". The idea is advanced that it is challenging to conceive of an adventure novel unfolding anywhere other than in the "human forest" of a sprawling metropolis. The more confined the setting (provincial town, neighborhood, or a single house), the more the adventure easily risks—according to the author's opinion—becoming internal, and the novel turning into an ordinary one.[2] (Pillat, 270) [*author's translation*]

[2] Original quote: "Al. A. Philippide, singurul scriitor român care s-a întrebat de ce romanul de aventuri nu a fost aproape deloc cultivat la noi, găsește cauza în lipsa."

Al. A. Philippide's perspective from "The Adventure Novel and Romanian Society" finds an advocate in Dinu Pillat, a connoisseur of sensational novel subgenres from the nineteenth century. This is because the inventiveness of Romanian writers seemed to be conditioned, on the one hand, by scenarios and strategies mediated through translations and, on the other hand, by their ability to adapt elements drawn from foreign cultural capital to the local specificities.

Fredric Jameson's assertion that "our daily life, our psychic experiences, our cultural languages are dominated by categories of space rather than by categories of time" encapsulates a core concept in his analysis of postmodernism and its cultural implications (Jameson 1990, 15). As postmodern scholars, emphasizing spatial considerations, we often engage in an analysis aimed at identifying spatial categories. Is it possible that the prevalence of spatial categories might be a distortion effected by the postmodern paradigm perspective? If this is the case, a focused examination of space-related concepts (applied to texts from any historical period) would, by logical extension, be considered a postmodern interpretation, aligning with Jameson's framework. This shift signals a broader cultural transformation wherein traditional linear narratives of time are supplanted by a focus on spatial configurations and relationships. The ascendancy of spatial categories implies a reorientation in our perception and navigation of the world. Everyday life, psychological experiences, and cultural expressions increasingly reflect the influence of spatial considerations, underscoring the importance of both physical and symbolic spaces in shaping our contemporary reality. This perspective aligns with the fragmentation and decentering characteristic of postmodernism, where an emphasis on space contributes to a sense of dislocation and disorientation. Moreover, in a society marked by rapid globalization, commodification, and the proliferation of images and signs, the spatial dimensions of culture become paramount. Jameson's observation underscores a fundamental shift in our cultural logic, indicating a departure from the primacy of temporal narratives to a terrain where spatial categories play a central role in shaping our experiences, identities, and cultural expressions.

While geocriticism and spatial criticism share a focus on the relationship between literature and space, they differ in their theoretical foundations and

methodological approaches. Geocriticism tends to emphasize the geographical imagination and the representation of landscapes, often using mapping techniques, while spatial criticism draws on broader theories of space from human geography and critical theory. Through the exploration of the contributions of influential theorists like Edward Said, Franco Moretti, Henri Lefebvre, and Michel de Certeau, literary studies can gain insights into the ways literature both influences and is influenced by the spatial dimensions of human experience. These methodologies enhance our understanding of the nuanced connection between textual and spatial elements in literary and cultural analysis, contributing to a deeper understanding of these complex dynamics. Geocriticism and spatial criticism concur on the notion that space not only generates the narrative elements but also is reciprocally reshaped by the narrative.

The "spatial turn" in literary studies

Both geocriticism and spatial criticism provide unique perspectives for examining the interconnection between space and literature. A relevant work in terms of understanding the transformative impact of spatial thinking across disciplines is Barney Warf and Santa Arias's *The Spatial Turn: Interdisciplinary Perspectives*. This work stands as a cornerstone in the examination of spatial thinking across disciplines. Published in 2009, the book captures the zeitgeist of a paradigm shift towards spatial considerations in literary studies. Its interdisciplinary approach not only encompasses geography and sociology but also extends to the realms of literary studies. The authors elucidate the theoretical foundations of the spatial turn, stressing the need to rethink conventional understandings of space and place (Warf and Arias 2009). This interdisciplinary perspective aligns with geocriticism, encouraging the blending of methodologies as a way to enrich analyses of literary spaces by drawing on insights from geography, anthropology, and cultural studies.

The geographical imagination, a key concept within geocriticism, aligns closely with Warf and Arias's call to reevaluate spatial paradigms. Other theorists, such as Franco Moretti, leverage mapping techniques to unravel spatial patterns in literature, providing a tangible link between the imaginative landscapes crafted by authors and

the geographical spaces they represent. Warf and Arias's emphasis on the multidisciplinary nature of the spatial turn resonates with the geocritical approach, which advocates for exploring literary spaces in comprehensive ways by integrating insights from various disciplines. Concurrently, spatial criticism, or literary cartography, draws inspiration from theorists like Lefebvre and de Certeau. Lefebvre's spatial triad and de Certeau's notions of strategies and tactics become invaluable tools for spatial critics exploring how literature reflects and challenges societal spatial practices (Lefebvre 1991; de Certeau 1984). By incorporating *The Spatial Turn* into the spatial criticism framework, research gains a more comprehensive understanding of how literature participates in the construction and contestation of spatial narratives.

Robert Tally's view of space as an active presence, which not only informs but also potentially determines various narrative elements, underscores the critical importance of spatial considerations in the realm of literary analysis. Tally's viewpoint aligns seamlessly with the theoretical framework of literary geography, accentuating the pivotal role that both the physical and symbolic dimensions of space play in shaping the profound substance of a narrative. The observation that space informs the narrative implies that the geographical and environmental context is fundamental to understanding the dynamics inherent in a story. Furthermore, Tally's claim that space may have significant impact on a narration underscores the likely that space will have a significant impact on the narration. This implies that geographic setting, cultural milieu, and physical surroundings may impose constraints or present opportunities that profoundly shape characters' actions and the development of the plot.

From raw data to a "Hajduk" literary space

The HAI-RO: *Hajduk Novels in Romania During the Long Nineteenth Century: Digital Edition and Corpus Analysis Assisted by Computational Tools* project emerged under the auspices of the "Brâncuși" Program for Integrated Actions, fostering a dynamic collaboration between the "Alexandru Ioan Cuza" University of

Iași and Université Paris 3 Sorbonne Nouvelle. Initiated by the project's principal investigators, Roxana Patraș and Ioana Galleron, the collaboration began as an experimental venture focusing on a series of TEI-XML novels for ELTeC (the "European Literary Text Collection" developed within the COST Action 16204: *Distant Reading for European Literary History*). This endeavor led to the formation of a dedicated French-Romanian team comprising Camelia Grădinaru, Ioana Lionte, Alexandra Olteanu, Lucreția Pascariu, Chiara Mainardi, and Ofra Lévy. In essence, the project addressed the deficiencies in resources and tools specifically tailored for Romanian Computational Literary Studies. Its research trajectory centered on characterizing the distinctive features inherent to "the Hajduk literature" by adapting corpus linguistics instruments to the nuances of literary texts printed in non-standardized Romanian. This endeavor was complemented by an examination of the external features developed throughout the historical evolution of a literary genre commonly categorized as "national." The primary outcome of the project is the creation of a Romanian literary corpus spanning from 1850 to 1950, meticulously aligned with the XML-TEI international standard of digital editing and enriched with semantic annotations. This corpus is accessible through an Open Access system, hosted on the Nakala platform (based in France), and documented on HAL-SHS.[3] This collection of Hajduk novels provides a rich source for exploring a dynamic literary space abundant in representations related to the motivations and circumstances behind characters' actions, sentimental embellishments of space, and the thematic patterns highlighted by scholars such as Tuan, Moretti, Warf, and Arias.

The HAI-RO collection constitutes a heritage corpus designed to delve into a literary tradition. The research initiative originated within genre theory, with a particular emphasis on examining the generic cohesion of the Hajduk subgenre. The primary aim was to scrutinize a phenomenon characterized by imitation and cultural

[3] HAL-SHS is an open access archive platform operated by the Centre pour la Communication Scientifique Directe (CCSD), a French national institution dedicated to providing free access to scholarly publications and research outputs in the humanities and social sciences. It serves as a repository for research materials, including articles, conference papers, and datasets, allowing researchers to share and access valuable academic resources worldwide.

transmission. This investigation was carried out through an analysis of the Hajduk hero's mobility; this analysis involved annotating spatial elements and examining narrative structures in mystery novels.

In the initial phase of mapping the literary landscape of the Hajduk novel, the HAI-RO project sought to establish a potential method for spatial annotation. This led to the identification of the following spatial categories: 1. places (*his village*, *poor wretched country*); 2. zones and regions (*the outskirts of Bucharest*, *in front of the house*); 3. imagined or archetypal spaces (topos) (*the other world*, *the universe*, *far from any village*); 4. elements related to an architectural plan (*at the church door*, *in the kitchen*); 5. spaces defined by body-related (e. g. clothing) elements (*near their breasts*, *from his pocket*, *on the head*); 6. transportation/access routes (streets, bridges); 7. vehicles (*in a carriage pulled by four horses*). In an effort to standardize the diverse spatial elements, a decision to extract toponyms was made, as a measure that increased the objectivity of the analysis and mitigated potential annotative ambiguities (Galleron *et al.* 2021). The optimal method for extracting toponyms was the use of UDPipeline (Straka *et al.* 2016). Dariah Geobrowser was used to generate a map of the Hajduk literary space.[4] UDPipe (short for *UD Parsing Pipeline*) is a natural language processing tool developed by Straka *et al.* in 2016. Dariah Geobrowser is a mapping tool used in this study to visualize the spatial data.

The identification and extraction of proper nouns hold significance within the realm of natural language processing (NLP), particularly in tasks such as information retrieval, named entity recognition (NER),[5] and sentiment analysis. The challenge presented by Hajduk novels, characterized by a diverse array of character names, locations, and named entities, underscores the relevance of using UDPipe for this

[4] I would like to express my heartfelt appreciation to my colleagues Ioana Lionte and Lucretia Pascariu for their dedicated efforts in collecting information about geolocations. Their commitment and diligence significantly contributed to the success of this research.

[5] Named Entity Recognition (NER) is a natural language processing task that involves identifying and categorizing entities within a text into predefined categories such as person names, organization names, locations, dates, and numerical expressions. NER is commonly used in information extraction, text mining, and other NLP applications to identify and extract important information from unstructured text data.

task. Developed by the Universal Dependencies project, UDPipe is an open-source library designed to process raw text across various languages. Its capabilities, including tokenization, lemmatization, syntactic parsing, part-of-speech tagging, and named entity recognition (NER), empower users to perform comprehensive linguistic analysis and information extraction tasks. While not specifically designed to extract toponyms (proper names) from texts, UDPipe can be used for this task, which involves leveraging the tool's tokenization and part-of-speech tagging capabilities for nouns and then generating CoNLL-U files.

Using UD Pipeline to extract toponyms and construct a map of the literary space within Hajduk novels involved multiple steps. Firstly, a careful selection of the novels was made, and these texts were then converted to .txt format and subsequently uploaded to the UDPipe system (*Figure 3*). The next step was tokenization, which involved breaking down the raw text into segmented units, known as tokens. Pre-trained models were used for analyzing the text and breaking it down into meaningful components. Part-of-speech (POS) tagging was then carried out, with UDPipe (as shown in *Figure 4*) used to assign grammatical categories (such as nouns, verbs, or adjectives) to each token. This process plays a crucial role in identifying proper nouns, which frequently denote names of characters, locations, or historical events. By employing techniques such as part-of-speech tagging, tokenization, and named entity recognition, researchers can effectively pinpoint these elements within the text. Preliminary results were obtained from the CoNLL-U file, from which all proper nouns were subsequently extracted (*Figure 4*).

The ensuing step involved manually curating nouns from the list of proper nouns. Once proper nouns that did not meet the criteria for toponyms had been eliminated, insights into the dynamic nature of the Hajduk subgenre unexpectedly came to light. The compilation of toponyms consisted of 597 entries, covering various types of Romanian settlements (villages, cities) as well as geographical regions with European, Asian, and American characteristics. This step revealed unexpected insights into the dynamic nature of the Hajduk novel.

The final list of toponyms was uploaded onto the Dariah Geobrowser, and any missing geographic coordinates were supplemented to ensure the precise location of

toponyms (*Figure 6*). The process of manually inputting geographical coordinates proves crucial to ensuring the quality of the data displayed on the prospective map, because this process contributes to eliminating ambiguities in those instances where toponyms are common to multiple geographical spaces, and to enhancing precision when it comes to locating both the sites where the Hajduks' adventures unfold and the spaces to which these sites refer.

The processes described above generated a map that illustrates the spatial features of Hajduk novels. UDPipe's adaptability is effective in extracting proper nouns from these novels, where efficient tokenization is critical to isolating individual words and facilitating analyses UDPipe is well-suited to the context-rich textual content of a Hajduk corpus. Proper noun extraction, which is an essential preprocessing step in various NLP applications, contributes to the understanding of cultural and historical narratives intertwined with the literary space.

The Hajduks' literary space

The map of the Hajduks' literary space is revealing not only in geographical terms but also thematically and historically. The most common toponyms are those located within the Romanian space, while those scattered throughout the world appear less frequently, despite being potentially meaningful in a regional novel subgenre (*Figure 7*). The frequency of toponyms that mark specific geographic areas delineates genuine Hajduk hubs clustered around the cities of Bucharest, Tecuci, Bacău, Pașcani, and aligning smoothly along the sub-Carpathian arch (*Figure 8*). Considering the infrequent travels of the Hajduks beyond the boundaries of the principalities, international toponyms do not function as active spaces for the development of intrigue; instead, they play a passive role within the characters' discourses, serving as venues for exploring ideas and for supporting the characters' impulse to reflect. The authors might have incorporated various international toponyms to evoke a broader cultural or geopolitical context, providing a sense of interconnectedness with the wider world. The inclusion of global toponyms could be a deliberate artistic choice,

allowing the author to create a more expansive and imaginative literary universe, transcending strict geographical boundaries.

The international toponyms might carry symbolic or allegorical significance, representing themes, ideas, or conflicts beyond the immediate physical locations, adding depth to the narrative. In *Figure 8*, which illustrates the distribution of Romanian toponyms, it is evident that assertions regarding the existence of Hajduk camps in mountainous regions are not confirmed. On the contrary, this hypothesis should be taken with a pinch of salt. Hajduk novels suggest that the mobility of rural outlaws is attributed not to rugged terrain but to the accessibility of commercial routes and postal roads. Certainly, the figure of the Hajduk would not have instilled such fear if he had not possessed the authority to regulate the movement of people and goods throughout the principalities of Wallachia and Moldova.

Juxtaposing *Figure 8*, which illustrates the literary realm of the Hajduks, with *Figure 9*, depicting the postal roads in Wallachia and Moldavia, reveals the Hajduks' familiarity with the pathways along which people and goods traveled within the principalities (*Figure 9*). The significant flow of traffic along these routes enabled Hajduks to use effective disguises, cross the territory inconspicuously, and gain quick access to information about potential targets and valuable assets. The rearrangement of Hajduk hubs in this revealing literary space reflects the political shifts of the latter half of the nineteenth century. The Hajduk novel underscores deep-seated connections with historical events, emphasizing the transfer of political power from Iași to Bucharest. Symbolically, Iași yields its authority, giving way to Bucharest.

Conclusions

The theoretical insights brought by geocriticism and spatial criticism methodologies, in conjunction with digital tools, provide a deep understanding of the Hajduks' literary space. The geographical choices of the authors are revealed through the strategic placement of toponyms on the map, and the extraction of toponyms was facilitated by the use of UDPipeline. This digital tool enabled the systematic extraction and analysis of toponyms from the text, contributing to the comprehensive

mapping of the Hajduks' literary space. The map of the Hajduks' literary space challenges assumptions about the rural outlaw camps in mountainous regions. The alignment of the Hajduks' literary territory with commercial routes and postal roads raises questions about the contrasting and conflicting dynamics between, on the one hand, subversive clandestine conduits represented by Hajduk roads and, on the other hand, the formal authorized networks embodied by postal routes. This alignment encourages reflection about the connection between rebellion and governance, as well as about the dichotomy between secrecy and transparency, elucidating the manifold roles that pathways play overtime. One may contemplate the intersections or disparities between Hajduk roads, often symbolizing resistance and opposition, and the meticulously controlled and regulated postal routes. Examining potential overlaps in specific regions prompts an exploration of what the coexistence or divergence of these pathways implies for the circulation of people, ideas, and information. This line of inquiry seeks to unearth the historical ramifications of such interactions, enriching our understanding of the complex historical milieu in which those pathways operated. The toponyms identified in the text shed light on the correlation between geographic spaces and characters in Hajduk novels.

The groups of toponyms are notably concentrated around Hajduk hubs such as Bucharest, Tecuci, Bacău, and Pașcani. This reveals the symbiotic relationship between the movement of goods and people along postal road maps in Wallachia and Moldavia and the tactics of the Hajduks to control not only their own mobility but also that of others. The intersection of the literary realm of the Hajduks with the maps detailing the circulation of goods and people encapsulates the characters' ability to exert authority over information flow, navigate territories inconspicuously, and execute rapid maneuvers. This exchange of political and symbolic power from Iași to Bucharest underscores the deep-seated connections between the Hajduk novel and the historical events it portrays.

A subsequent research project that examines Hajduk novels, while expanding the spatial perspective to incorporate verb analysis, has the potential to reveal interesting connections between various physical configurations such as architecture, landscapes, roads, and geographical features, alongside human emotions and

narrative depictions, and narrative depictions. Drawing on Yi-Fu Tuan's perceptive insights into the kinesthetic interplay between emotions and physical surroundings, the exploration of the verbs used to portray Hajduk landscapes offers a rich avenue for investigation. Through meticulous analysis of the dynamic language used in these novels, we can observe the implicit ways in which emotions and perceptions associated with the depicted spaces are conveyed. For instance, verbs characterizing mountain peaks as "soaring" or ocean waves as "swelling" not only mirror the physical attributes of the landscapes but also evoke a sense of grandeur or power. Furthermore, this study may consider the impact that architectural forms have on perceptions of space—as suggested by Tuan—by analyzing verbs that delineate spatial expansion and contraction within Hajduk narratives. Such an inquiry promises to enrich our understanding of the interaction between language, spatial representation, and emotional resonance within the realm of Hajduk novels.

The Hajduk novels exhibit a unique conceptualization of spatial dynamics, distinguishing between active and passive spaces. Active spaces go beyond their traditional role as mere settings, influencing the narrative trajectory. These dynamic arenas serve almost as theatrical stages where Hajduk protagonists navigate, resist, and orchestrate their actions and challenges. In contrast, passive spaces dispersed globally play a crucial role in serving a distinct imaginative function within the Hajduk discourse. They predominantly act as contemplative venues that facilitate exploration of broader cultural or geopolitical contexts.The inclusion of global toponyms in the novels can be interpreted as an artistic decision that fosters the creation of an expansive and imaginative literary realm, enriching the narrative with broader cultural or geopolitical contexts. This dichotomy between active and passive spaces not only enhances the complexity of the Hajduk narrative but also encourages readers to delve into the symbolic, thematic, and geopolitical dimensions integrated into the spatial configurations of these novels. The Hajduk literary space is an intriguing construct, providing readers with a distinctive perspective from which to examine the intersection of literature, space, and history.

***Annex 1*: Figures**



Figure 1: Sample1- illustrated page from a Hajduk novel



Figure 2: Sample2- illustrated page from a Hajduk novel

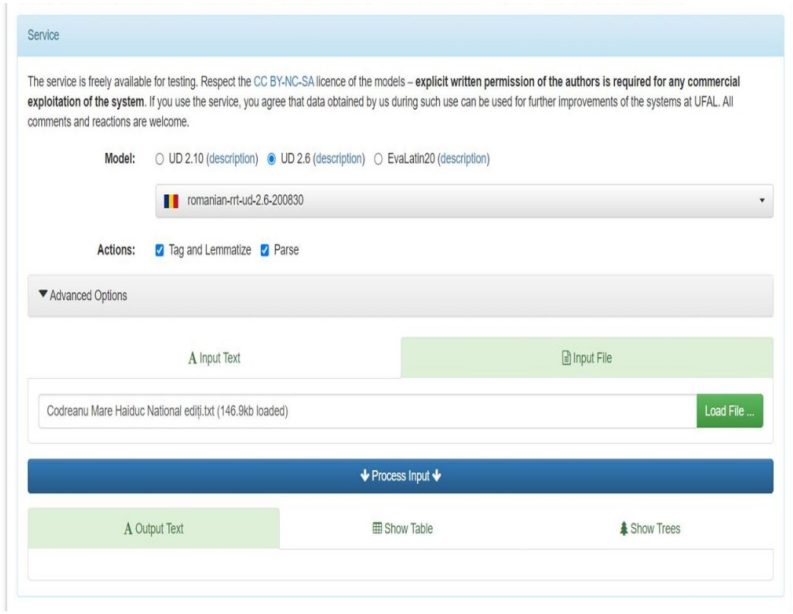


Figure 3: Loading a Hajduk novel as an input file

Save Output File

Id	Form	Lemma	UPosTag	XPosTag	Feats	Head	DepRel	Deps	Misc
# generator = UDPipe 2, https://lindat.mff.cuni.cz/services/udpipe									
# udpipe_model = romanian-rrt-ud-2.10-220711									
# udpipe_model_licence = CC BY-NC-SA									
# newdoc									
# newpar									
# sent_id = 1									
# text = Moldova în veacul al 17-lea Istoria ce vom povesti în această broşură, se petrece în a doua jumătate a veacului al şépte-spre-decelea, cam pe la anul 1666, Moldovei în timpul acela, pentru ca faptele ce vom povesti să nu pară nici stranii nici imposibile.									
1	Moldova	Moldova	PROPN	Np	_	15	nsubj	_	TokenRange=0:7
2	în	în	ADP	Spsa	AdpType=Prep|Case=Acc	3	case	_	TokenRange=8:10
3	veacul	veac	NOUN	Ncmsry	Case=Acc,Nom|Definite=Def|Gender=Masc|Number=Sing	15	obl	_	TokenRange=11:17
4	al	al	DET	Tsms	Gender=Masc|Number=Sing|Poss=Yes|PronType=Prs	5	det	_	TokenRange=18:20
5	17-lea	17-lea	NUM	Mc-p-d	Gender=Masc|Number=Sing|NumForm=Word|NumType=Ord	3	nummod	_	SpacesAfter=\r\n|TokenRange=21:27
6	Istoria	istorie	NOUN	Ncfsry	Case=Acc,Nom|Definite=Def|Gender=Fem|Number=Sing	15	nsubj	_	TokenRange=29:36
7	ce	ce	PRON	Pw3--r	Case=Acc,Nom|Person=3|	9	obj	_	TokenRange=37:39

Figure 4: Results of POS tagging as an output file

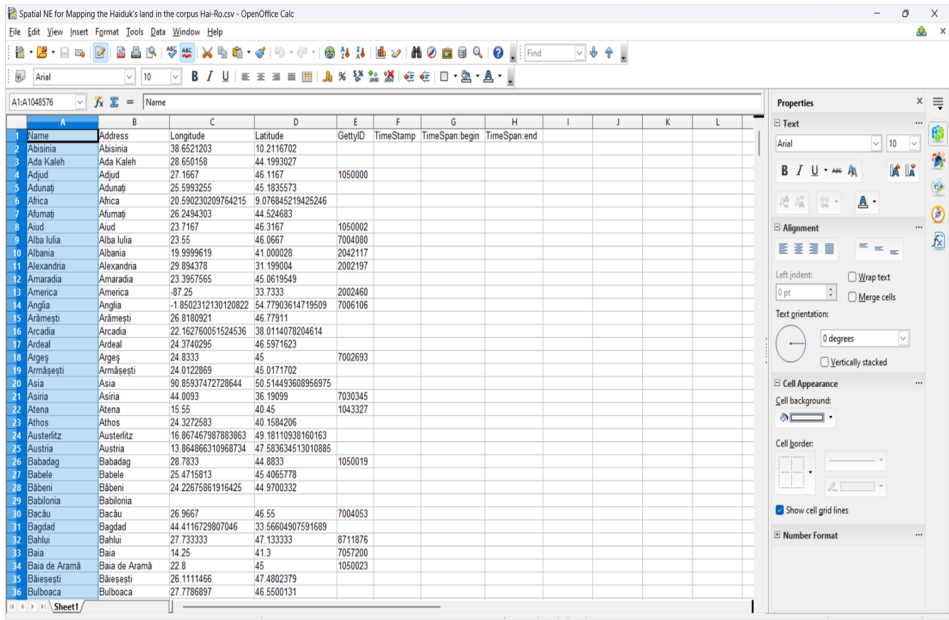


Figure 5: The final list of 597 toponyms

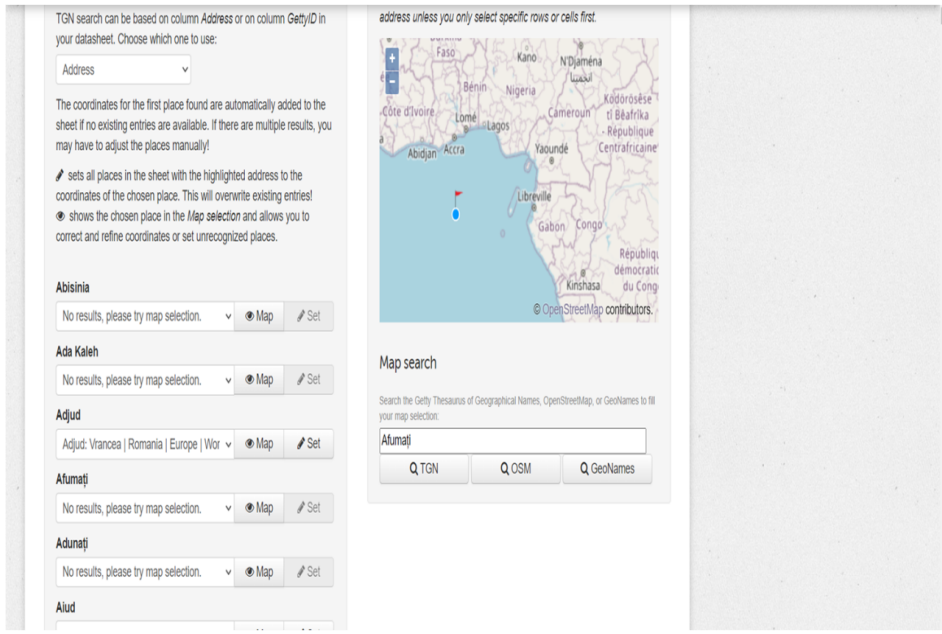


Figure 6: The manual addition of geographical coordinates using Geobrowser



Figure 7: The world seen through Hajduks' words

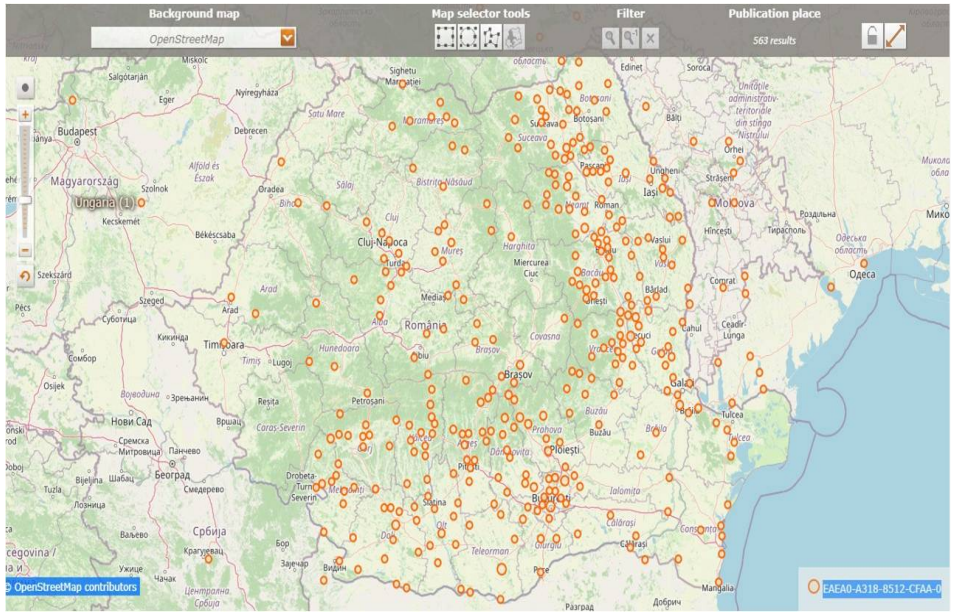


Figure 8: The crystallization of Hajduk hubs

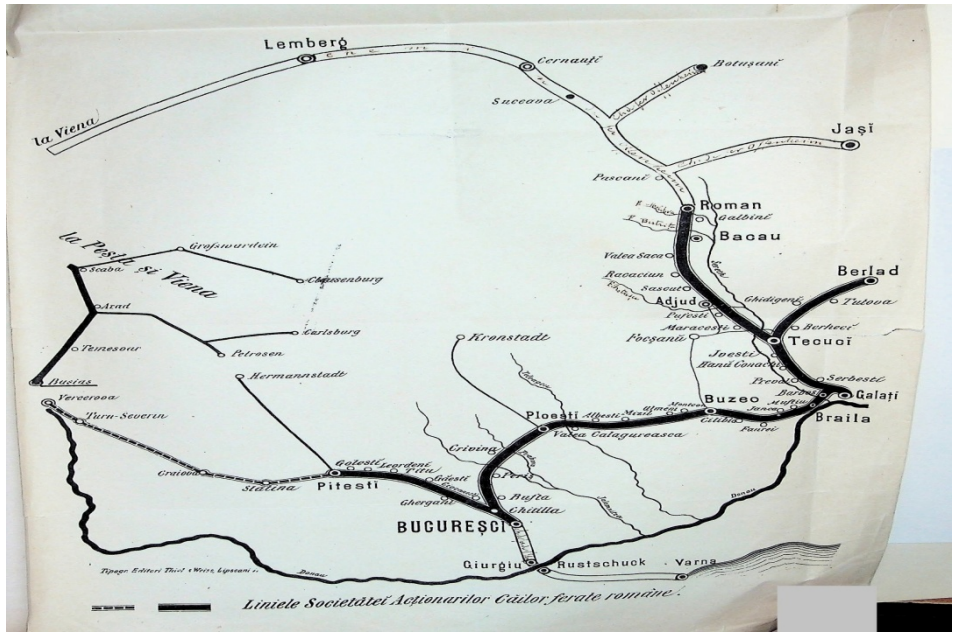


Figure 9: The map of postal roads in Wallachia

References

De Certeau, Michel. 1984. *The Practice of Everyday Life*. Berkeley: University of California Press.

Galleron, Ioana, Roxana Patraș, Camelia Grădinaru, and Frédérique Mélanie-Becquet. 2021. "À la recherche des haïdouks. Annoter les entités spatiales dans des romans roumains du XIXe siècle/Looking for Hajduks. Space Entities Annotation in Nineteenth-Century Romanian Novels." *Humanités numériques* 3 (3). https://doi.org/10.4000/revuehn.1399.

Jameson, Fredric. 1990. *Postmodernism, or the Cultural Logic of Late Capitalism*. Durham: Duke University Press.

Kennedy, Dane. 1996. "Imperial History and Postcolonial Theory." *The Journal of Imperial and Commonwealth History* 24 (3): 345–363.

Lefebvre, Henri. 1991. *The Production of Space*. New Jersey: Wiley-Blackwell.

Moretti, Franco. 1999. *Atlas of the European Novel 1800–1900*. London: Verso.

Patraș, Roxana, Ioana Galleron, Lucreția Pascariu, Alexandra Olteanu, Ioana Lionte, and Camelia Grădinaru. 2020. *HAIRO: Corpus de Romane Haiduceşti (1850–1950)*. https://www.nakala.fr/page/collection/11280/51a881d1.

Patraș, Roxana, Lucreția Pascariu, Alexandra Olteanu. 2023. "Patrimoniu și Inovație. HAI-RO/ELTeC/Pop-Lite." Paper presented at the SmartDiaspora conference, Timișoara.

Said, Edward. 1994. *Culture and Imperialism*. New York: Vintage.

Straka, Milan, Jan Hajič, and Jana Straková. 2016. "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing." In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, May 2016.

Tally, Robert T. (*ed*). 2014. *Literary Geographies: Spatiality, Representation and Narrative*. New York: Palgrave MacMillan.

Tuan, Yi-Fu. 1990. *Topophilia: A Study of Environmental Perception, Attitudes, and Values*. New York: Columbia University Press.

Warf, Barney, and Santa Arias. 2009. *The Spatial Turn: Interdisciplinary Perspectives*. London: Routledge.

Westphal, Bertrand. 2011. *Geocriticism: Real and Fictional Spaces*. Translated by Robert T. Tally. New York: Palgrave MacMillan.

DIGITUS DEI EST HIC!

Proxy Structures in Computational Analyses of Text Corpora

Laura PRICOP

Digitus Dei est Hic!

Since the turn of the twenty first century, humanity has been cautiously entering the digital realm, impacting culture, society, economics, and human well-being. This is a realm in which distance as much as data is compressed, and feedback is instantaneous. Are we prepared for this new world? Or, perhaps more accurately: How are we preparing ourselves for this world?

Whether engaged in learning, working, or building relationships with family members or romantic partners, we conduct at least half of our lives through computer screens. Despite the hectic rhythm and apparent chaos of transitioning from the pre-digital era, there seems to be a sort of control, better yet, a controlled desire to achieve, in *compositionist* terms, a "tentative and precautionary progression" (Latour 2010). The current European or national legislation (horizontal or sectoral), enables different institutions and sectors of society—such as government, economy, transport, education, health, sport, culture and cultural heritage—to make their services available within digital environments. The scientific field of humanities, are undergoing a similar process of digitization, which inspired Roberto A. Busa's famous declaration: *Digitus Dei est hic!* [the finger of God is here!] (Busa 2008). Equally, the presence of artificial intelligence (AI) in our everyday lives is no longer being called into question because AI recently became the engine of the mass digitization phenomenon across application fields. Experts define AI as a cluster of systems of software technologies (e.g. voice assistants, image analysis software, search engines, voice and facial recognition systems) or hardware (e.g. advanced

robots, autonomous vehicles, drones or Internet of Things applications) designed by humans to accomplish complex goals in the physical or digital dimension. These systems have the ability to sense the environment by collecting unstructured data, to interpret knowledge or process information from this data, and then to decide on the best course of action to be taken to achieve a given goal. AI systems can also adapt their behavior by first analyzing how their previous actions have affected the environment and then integrating new symbolic rules or learning numerical models. Recently, AI specialists are focusing on the positive or negative quality of AI usage and on the magnitude of the effects it can have: "the pertinent questions now are by whom, how, where, and when this positive or negative impact will be felt" (Floridi *et al.* 2023).

Therefore, with a view to AI usage the human condition and human dignity in the digital age, scholars in the area had to reformulate several philosophical questions:

- Who can we become? In other words, to what extent is autonomous self-realization possible if it is conditioned externally by technology?
- What can we do? How much freedom do human beings still have if decision-making is initiated and verified by technology?
- What can we achieve? How can individual and societal capabilities still be boosted if such activities are partially undertaken by technology?
- How can we interact with each other and the world? How might societal cohesion still be realized and sustained if it is mediated by technology (Floridi *et al.* 2023)?

My view is that these questions will become even more frequent over time and that, with the "controlled" progress of knowledge, their answers will only ever be partial and consist of temporary responses to punctual issues.

Providing a broad oveview of the main questions being asked, this chapter encourages an ethical approach to studying how the use of computational tools affects human thinking habits and behaviour. In this chapter, I attempt to analyze the behavior of the humanities, when they are engaged in processes of digital transformation. Epistemologically, I am interested in hybridization of literature as a

field of knowledge; in terms of logic, I am interested in how proxy arguments, work within the questions posed on computing texts. Proxy arguments refer to an argument made on behalf of or in place of another argument to support a main argument, adress weaknessess, and bringing in supplimnetary evidence.

Compatibility between Logic and Literature

Logic and literature are two areas of knowledge that have commonalities in their development. If Plato expelled poetry/literature from the city, Aristotle classified entities in the natural world and categorized various types of utterances. The *Organon* not only showcased Aristotle's analytical prowess but also established a robust correlation between language analysis and the classification of utterances. By delving deep into the nuances of linguistic structures and patterns, the Greek philosopher was able to uncover the underlying connections between language and the different ways in which people express themselves. His groundbreaking work not only enhanced our understanding of language but also shed light on the intricate relationship between logic and human communication. For several centuries, logic was placed, in Aristotle's tradition, alongside grammar and rhetoric as *ars sermocinalis*, and all of the arts of discourse (seen as the means by which people communicate with each other), including literary theory, were later classified in this category (Kneale and Kneale 1975, 377; Murphy 1974, 91).

Recent interdisciplinary studies have attempted to develop a new field in which logic and literature engage in philosophical dialogue.[1] Scholars argue for the "trans-historical portability of literary form to historical accounts of logic and logical form" (Belvins 2020). Accordingly, logic and literature should be devoted both to exploring the possibilities of reconciling logical rationality with the inherently imaginative and

[1] The concept of interdisciplinarity still provokes much debate. Extensive discussion of these arguments is beyond the scope of this chapter; I would, however, like to highlight that Wolf defines interdisciplinarity as a philosophy of integrative thinking (Wolf 1998). By "interdisciplinary philosophy," I mean the complementary nature of philosophy, which provides a solid foundation and rigorous tools for integrating—and perhaps understanding in more logical ways—the results of research meticulously undertaken by digital humanities specialists.

affective aspects of literature, and to comparing narrative analysis with logical systems of inference. Grounded in this "trans-historical portability," the compatibility between literature and logic may be addressed within a narrow interdisciplinarity framework. Still, when we consider their comparability within the framework of digital humanities, a field that reflects hybridity and experimentalism *par excellence*, then a higher degree of interdisciplinary integration—in formal as well as methodological terms—may be achieved (Newell 1998). Borrowing a method or concept from one discipline to use it in another serves the aim of testing hypotheses, developing new theories, and/or responding to problems that could not be solved with the tools available within the researcher's immediate disciplinary paradigm. With the advent of distant reading, literature seems to have internalized quantitative methodologies: this has led to more fine-grained interpretations of textual phenomena but also to the concern that the resulting findings are not robust enough. In fact, computational analyses draw from traditional philological practices, such as indexing, enumeration of forms, and concordance—each of which the computer can perform more quickly if the texts being processing meet certain criteria, like addressability (Wittmore 2016). However, big data also undergoes a process of abstraction or reduction so that it can be easily analyzed and used for modelling.

It seems reasonable to assume that technology-based quantitative methods used in the empirical sciences could be transferred to literary studies. It is possible to adapt digital tools to analyses of literary content if we draw on both a larger interdisciplinary paradigm and basic logic It is due to this interdisciplinarity that proxy argument is used. In short, in computational text analysis, proxy argument seems to play the role of stabilizing relations between terms or sentences through analogy.

This chapter, which generally falls within the realm of interdisciplinary studies (or rather interdisciplinary philosophy), aims to clarify the structure of proxy argumentation and how it is used in the digital humanities. Although digital humanities research relies on the adaptation of proxy measures and arguments for data selection and interpretation, this knowledge behavior has only been partially theorized. What are these measures and arguments? How are they constructed? How

can they be justified? And how can their validity be verified? In this chapter, I will begin to think about these questions—which are as yet unexplored—by examining the complex network of associations that the proxy argument unveils, enriching the dynamic interaction among different sentences, entities, and attributes conceived within fictional realms. By doing so, I will provide fresh insights into the application of logic in the field of digital humanities, as the application of logic establishes novel points of connection between fresh insights into how concepts from the field of logic can be applied to the field of digital humanities. This exploration will illuminate the ways in which logic is harnessed to navigate the complexities of digital texts and their spatial representations, ultimately enriching our comprehension of both disciplines.

Structure and functions of the argument by proxy

Within the field of logic, literary texts have been disputed for their inability to trigger logical inquiry and have been accused of not containing a sufficient amount of argumentative material (Kneale and Kneale 1974, 9). However, literature, theatre, film and online videogames that take place in virtual environments undeniably describe or enact complex situations that are grounded in moral reasoning (Søndergaard 2021), a basic type of argumentative structure that has often been discussed in relation to elliptical reasoning. Literature, in particular, seems to contribute to the accumulation of knowledge, and this knowledge seems to raise claims to validity by virtue of the moral change, it produces.

From a logical point of view, computational analyses show the tendency to establish binary proxy relations between simple propositions in order to support the epistemological thesis under consideration. A binary proxy relation can take the following form (Fritz 2019):

"Xa" ("**a** to this **X**")

"Xa" can be "Yb"

The conditions for Xa to be Yb are not restricted to the equivalence between a and b or between X and Y. Thus, Xa can be Yb without a being b or X being Y. The

relation a to X is structured by proxy by the proposition that relates a to Y. The use of by-proxy argument schemes does not generate all true sentence conditions (Nelson 1997). True sentence conditions relate to evidence that is empirically verifiable. This shortcoming in the digital humanities can be compensated for by ways of evaluating arguments and an open attitude to how truth is delivered (Symons 2017, 31).

Let us consider a paradigmatic case of an example in natural language: "Raindrops are colorless if and only if tears are colorless" (Xa if and only if Yb). We can see that the statement "tears are colorless" does not necessarily fulfil the truth condition. Even though the premise of tears being composed of blood may seem fictitious, hemolacria is a real medical condition that disrupts the truth value of "tears are colorless." In this case, the truth of the statement "raindrops are colorless"[2] cannot be obtained by the proposed proxy argument. On the other hand, the Romanian expression *apă de ploaie* [rainwater]—referring to superficiality—raises truth claims if it has as a proposed proxy argument the Romanian expression *lacrimi de crocodil* [crocodile tears]. This is valid in the case of the Romanian language, so the proposed analogy has a restricted degree of validity due to the context of use. If the expression "rainwater" (Rw) is a proxy for superficiality, then the expression "crocodile tears" (Ct) can support the idea of a behavioral manifestation affected by superficiality. If Rw is a proxy for lack of content, then Ct loses its function of supporting Rw, and the analogy between the two expressions is false; it is a faulty analogy (Govier 2014, Symons 2017, 163-178). There are other types of errors related to the use of analogy: the two-wrongs fallacy, the fallacy of slippery uses of assimilation, and the fallacy of slippery uses of precedent. The two-wrongs fallacy is based on the two-wrongs analogy, whereby, if two wrongs are similar and one of them is tolerated, then the other, by virtue of consistency, should also be tolerated. The fallacy of slippery assimilation is a false appeal to consistency that starts from the premise that, if the differences between successive situations are small, they must all be accepted. This fallacy actually cancels out the small differences between entities which, if

[2] The truth value of this statement changes if one takes into account the rare meteorological phenomenon that produces the so-called "bloody rain" caused by the presence of Saharan dust particles.

aggregated, would prove significant. The fallacy of slippery precedent argues that a good deed should not be allowed, because this will create a precedent for committing other deeds that are bad. It misreads the distinction between right and wrong.

Some analogies can be convincing, and others can be irrelevant or misleading. In general, two types of analogies are accepted: *a priori* and inductive. *A priori* analogies depend on consistency and support the resolution (acceptance or refutation) of conceptual situations. The analogy may be either hypothetical or fictitious, but its value does not alter the contextual understanding. Inductive analogies have an important epistemological function because they relate known structures to new, unexplored ones. In computational analyses, proxy structures based on *a priori* analogies are frequently found in lexicon-based studies, whereas inductive ones are adopted in analyses that favor experiment and allocate more freedom in terms of both establishing the methodology and exploring and interpreting the results obtained.

Let us now return to the structure of the argument by proxy. If the truth claim is dropped in the case of Yb in order to assert Xa, then it is recommended (Nelson 1997) that the by-proxy argument be combined with Quine's observation theory, because it is only with this conceptual support that the ontological relativity of a proposition can become an epistemological thesis. Quine's observational propositions are the starting point in our direct acquisition of sense-based knowledge. For an observational statement to be accepted, the condition of perceptual similarity must be met. A sentence such as "It is raining outside" can be accepted or rejected if observers receive sensory stimuli that support or contradict the proposition (Quine 1969).

If Yb is accepted as a proxy argument to support Xa, then further explanations are needed to justify the acceptance of the argument. Logic imposes that the acceptability of grounds should be exclusively related to the content of the argument and not to the argumentative structure. In other words, an argument may contain one or more premises that are presented as evidence to support the truth of the conclusion. According to Govier's list, the conditions for the acceptability of a premise in an argument may be as follows:

a) the premise is supported by a cogent sub-argument;

b) the premise is supported elsewhere by the arguer or another person, and this fact is noted;
c) the premise is known *a priori* to be true;
d) the premise is a matter of common knowledge;
e) the premise is supported by appropriate testimony—that is, the claim is not implausible, the sources are reliable, and the content of the claim is restricted to experience.

As far as the argument's content is concerned, explanation is based on a different balance of acceptability (Govier 2014, 128). The *explanandum* (i.e. what has been explained) is assumed to be true, and the explanatory factors can also be assumed to be true and used to elucidate causal phenomena, narratives, motives or intentions. Explanations are not separate from argumentation. They can, as Govier mentions,[3] complement and support each other.

In computational analyses, it is essential that by-proxy structures be supported by explanations. Otherwise, the methodological approach is opaque and the final result may not be deemed reliable. Disregarding from the truth conditions of a proxy argument shifts the problem of truth over to the realm of linguistics and, more specifically the interpretation of meaning.[4] If we start from the simple fact that computational analyses usually process corpora of texts (that is, a set of word units and their contexts), then we might incorrectly assume that the meaning of an isolated utterance is equal to that utterance's deeper meaning (Meyer 1986, 284–285). This situation also indicates that an utterance becomes more informative if reattached to its context; therefore, the meanings of particular utterances become visible as part of the contexts. For instance, in the case of temporal mappings, it has been observed that

[3] Among the arguments supported by explanations are abductive arguments (arguments that support the best explanation for a given phenomenon) and transcendental arguments (arguments that require that between A and B there is a necessary condition that can be explained if A is used as a factor that explains B).

[4] The definition of "meaning" and the distinction between *meaning* and *reference* are among the issues that are still hotly debated today. Quine stated that "pending a satisfactory explanation of the notion of meaning, linguists in semantic fields are in the situation of not knowing what they are talking about". The concept of *meaning* in Quine's terms is a special kind of entity, and the meaning of an expression is the idea expressed (Quine 1980, 47-48).

"relative TOs (temporal operators) are usually intuited from the context of story, rather than directly stated" (Yager 2023).

Many computational analyses are grounded in by-proxy arguments. In digital humanities, variations depend on previous practice, on tools and on the complexity of the corpora under study. To support the by-proxy argument, computational analysts of text corpora may resort to developing complex annotation schemes, as in the case of spatial entities proposed for a Romanian corpus of 19th-century hajduk novels. In practice, researchers have identified and classified toponyms, common place names or locative prepositions specific to the Romanian language. Therefore, in the first phase of the study, in which observation units were initially identified, morphologic categories that carried spatial information functioned as proxies. Then, in order to highlight orientation, movement and metrical spatial relations, researchers expanded the sample to include more complex syntactic units and "focused on nouns and adverbs carrying spatial semantics, taking into account the whole nominal, pronominal or adverbial phrase" (Galleron *et al.* 2021). The spatial turn is also a concern of Grisot and Herrmann (2023), who aim "to detect and examine how different types of space are distributed and affectively encoded in German-Swiss literature." In this case, the researchers detected *emotion* or *mood words* in a corpus of 125 texts written in German by establishing two broad spatial categories: rural, urban. The creation of emotion lexicons related to spatial entities (in groups of approximately 50 words clustered around the detected entities) served as a proxy for mapping rural and urban spaces as sentimental landscapes rather than as architectural or social configurations. As the authors note, these lexicon-based approaches (recently, the term "bag of words approach" has been used self-critically or pejoratively) faced a number of challenges such as those related to the finite number of words included in the compiled list, to the inconsistency between different lexicons and a certain definition of a concept, and to the difficulty of representing exceptional situations in a ready-made scheme and of devising new functional categories to account for unexpected data encountered during in the research process

Discussion and conclusion

The recent development of digital humanities has brought with it harsh criticism. The field is unfairly accused of, among other things, betraying principles of humanities and of distancing itself from traditional academic research and pedagogy, the main concern being that students will receive training in the use of technologies rather than in critical thinking (Galleron and Imhadad 2020). However, the appropriateness of criticism to reveal, to debunk prejudices, to enlighten and encourage minds in the shattering of illusions and idols is also questioned by Bruno Latour. He outlines a way of reusing criticism, "a critique acquired secondhand—so to speak—and put to a different use" (Latour 2010). The idea of composition on which Latour's process of reusing criticism is based concerns the economy of knowledge (Ciubotarașu-Pricop 2014), which involves combining 1) knowledge that has been acquired through experience and 2) processed experience, such as inferred effects with respect to the usability of an object or phenomenon. Descartes established three different modes of composition: by impulse, by conjecture and by deduction. In the case of composition by impulse, one makes decisions based on belief in a higher power, based on one's own freedom or based on one's imagination. Judgments derived from reached through imagination are credited with the lowest confidence value, and these are followed by judgments obtained by conjecture, which according to Descartes "do not make us more learned" but have a higher degree of probability. In the case of deductive judgments, they are beyond doubt but not beyond error, as in the deduction of a general and necessary thing from a particular or contingent one (Descartes 1964). A follower of Descartes and Locke, Bruno Latour's compositionism—proposed as an alternative to criticism—aims to reassemble reality piece by piece. For a compositionist, everything is called into question, judgements are touched by Cartesian doubt, nothing is revealed by criticism, science (with a small "s") is achieved "only by the slow process of composition and compromise, not by the revelation of the world beyond" (Latour 2010).

By recovering imagined, fictional or real worlds, digital humanities scholars carry out compositional labor that needs to be discussed in appropriate terms. The

challenge lies in adapting to language processing a methodology that was developed within the sphere of empirical sciences. Unlike classical logic, textual analysis requires the use of logical forms that closely align with natural/intuitive thought processes. Therefore, analytical processes have shifted from traditional logical reasoning and introduced new approaches such as the argument by proxy.

The methods of the compositional project may establish a connection between two different propositions, supported by explanation and analogy. By using argumentation by proxy, researchers are able to operationalize categories. Knowledge acquired through this process serves to compensate for discontinuities within the text. It also helps to establish and locate new relationships between different observation units. By employing logical forms that align with natural thought processes on the trial-and-error paradigm, digital humanities researchers are able to uncover hidden connections within texts and shed light on the intricate complexities of textual collections. This painstakingly constructed new knowledge makes use of creativity and empathy but also requires rationality to process large datasets.

Finally, what is the special contribution of quantitative analyses? One of the most elegant answers to this question can be found in Witmore's suggestion that computational analysis has the ability to capture the finest and most imperceptible details in a text, as in the full collection of Eadweard Muybridge's photographs capturing the galloping horse: the observer can see the stride with all four feet in the air, unless their perception is mediated by the sequence of snapshots (Witmore 2016). The snapshots of the horse represent a movement that, unless recorded and played back by a camera, is not accessible to the naked eye. The same movement of recording discrete units and playing them back all over again through trial-and-error iterations seems to be the distinctive feature of research in digital humanities.

References

Belvins, J., and D. Williams. 2020. "Logic and Literary Form." *Poetics Today* 41 (1). https://doi.org/10.125/03335372-7974058.

Botezatu, P. 1974. *Semiotics and Negation: Critical Orientation in Modern Logic*. Iași: Junimea Publishing House.

Busa, A. R. 2008. "Foreword: Perspectives on the Digital Humanities." In *A Companion to Digital Humanities*, edited by S. Schreibman, R. Siemens, and J. Unsworth, 10–15. Blackwell Publishing.

Ciubotarașu-Pricop, L. 2014. *Problematologie și discursivitate. Cu un exercițiu asupra pasiunilor*. Iași: Alexandru Ioan Cuza University Publishing House.

Descartes. 1964. *Reguli utile și clare pentru îndrumarea minții în cercetarea adevărului* [Discourse Method of Correctly Conducting One's Reason and Seeking Truth in the Science]. Translated by Cornelius Vilt. Bucharest: Editura Științifică.

Floridi, L., J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena. 2018. "AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations." *Minds & Machines* 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5.

Fritz, P. 2019. "Structure by Proxy, with an Application to Grounding." *Synthese*. https://doi.org/10.1007/s11229-019-02450-z.

Galleron, I., R. Patraș, C. Grădinaru, and F. Melanie-Becquet. 2021. "In Search of Haidouks. Annotating Spatial Entities in 19th Century Romanian Novels." *Humanites Numeriques* 3. https://doi.org/10.4000/revuehn.1399.

Galleron, I., and F. Idmhand. 2020. "Why Go from Texts to Data, or The Digital Humanities as a Critique of the Humanities." *Word and Text: A Journal of Literary Studies and Linguistics* X: 53–66. halshs-03951355.

Govier, T. 2014. *A Practical Study of Argument*. Enhanced Seventh Edition. Wadsworth: Cengage Learning.

Grisot, G., and B. Herrmann. 2023. "Examining the Representation of Landscape and Its Emotional Value in German-Swiss Fiction between 1840–1940." *Journal of Cultural Analytics* 8 (1). https://doi.org/10.22148/001c.84475.

Haverals, W., L. Geybels, and V. Joosen. 2022. "A Style for Every Age: A Stylometric Inquiry into Crosswriters for Children, Adolescents and Adults." *Language and Literature* 31 (1): 62–84.

Kneale, W., and M. Kneale. 1974. *Dezvoltări ale Logicii* [The Development of Logic], vol. I. Translated by Cornel Popa. Cluj-Napoca: Editura Dacia.

Kneale, W., and M. Kneale. 1975. *Dezvoltări ale Logicii* [The Development of Logic], vol. II. Translated by Sorin Vieru and Uşer Morgenstern. Cluj-Napoca: Editura Dacia.

Latour, B. 2010. "An Attempt at a *Compositionist Manifesto*." *New Literary History* 41: 471–490.

Meyer, M. 1986. *From Problematology: Philosophy, Science, and Language*. Brussels: Pierre Mardaga.

Murphy, J. J. 1974. *Rhetoric in the Middle Ages: A History of Rhetorical Theory from Saint Augustine to the Renaissance*. Berkeley: University of California Press.

Nelson, R. J. 1997. "Proxy Functions, Truth and Reference." *Synthese* 111: 73–95.

Newell, W. H. 1998. "Professionalizing Interdisciplinarity: Literature Review and Research Agenda." In *Interdisciplinarity: Essays from the Literature*, edited by W. H. Newell. New York: College Entrance Examination Board.

Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. 2021. COM/2021/206 final, April 21. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PCO206.

Radak, T., *et al.* 2023. "Toward a Computational History of Modernism in European Literary History: Mapping the Inner Lives of Characters in the European Novel (1840–1920)." *Open Research Europe*. https://doi.org/10.12688/openreseurope.16290.1.

Quine, W. V. 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W. V. 1980. *From a Logical Point of View: 9 Logico-Philosophical Essays*. Harvard University Press.

Schreibman, S., R. Siemens, and J. Unsworth (eds). 2008. *A Companion to Digital Humanities*. Oxford: Blackwell Publishing.

Søndergaard Christensen, A. M. 2021. *Moral Philosophy and Moral Life*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198866695.003.0007.

Symons, J. 2017. *Formal Reasoning: A Guide to Critical Thinking*. Dubuque: Kendall Hunt Publishing Company.

Witmore, M. 2016. "Latour, The Digital Humanities, and the Divided Kingdom of Knowledge." *New Literary History* 47 (2–3): 353–375. https://doi.org/10.1353/nlh.2016.0018.

Wolf, S. 1998. "The Challenge and Satisfaction of Integrative Thinking." *Integrative Physiological and Behavioral Science* 33 (2): 120–121.

Yager, S. A. 2023. "Time Maps: Theory and Methods." *Journal of Cultural Analytics* 8 (3). https://doi.org/10.22148/001c87937.

The Many Faces of *Virtual*: charting the family resemblance network for conceptual understanding

Matei Alexandru STOENESCU

Semantic fuzziness and a full spectrum of meanings for *virtual*

The term *virtual* and its related concepts have become ubiquitous across modern scientific discourse and cultural narratives. This family of concepts, dynamic and expansive, has the potential to reshape our cognitive and behavioral frameworks across multiple fields. *Virtual* carries a broad array of interpretations—some sharply delineated, others more fluidly connected. This study sets out on an epistemological quest to delineate and elucidate this cluster of meanings through the lens of family resemblances, as originally conceptualized by Ludwig Wittgenstein in his seminal work, *Philosophical Investigations*.

This article builds upon our prior research, concluding with the Ph.D. thesis *From Fictional Worlds to Virtual Worlds. Drivers of Change Between Perspectives on Fictional, Virtual, and Real.* In this previous work, we explored the essence of fictional and virtual worlds through the lens of the possible-worlds theory, principally as delineated by David Lewis (Lewis 2). We showed that fictional and virtual worlds constitute a dynamic framework intimately connected to the real world, suggesting that each entity regarded as *fictional*, *virtual*, or *real* can transition into one of the other states. Furthermore, we identified and discussed what we believe to be the fundamental drivers of these transformations. As part of the research process, we endeavored to chart the array of meanings attributed to the concept of the virtual, aiming to reveal the complex interconnections that form its family resemblance network. To that end, we gathered all the semantic nuances of "virtual" found in the

referenced literature, along with the term's diverse uses. We then manually plotted these findings on a dynamic Kumu chart, carefully delineating the relationships between uses and features. This endeavor was marked by three essential qualities: a high degree of accuracy in aligning uses with features, a considerably rich set of *semantic features*, and a relatively small initial set of uses, which was a direct consequence of the meticulous manual mapping process we undertook.

Despite the precision of our manual mapping and identifying the sets of uses and semantic traits, we must address the concern that this process does not entirely reflect the natural use of the term "virtual" in natural language. Hence, this article presents an alternative method aiming to capture a broader and more objective set of "virtuality" uses. Furthermore, in contrast to the previous dynamic chart—which primarily highlighted the intricate relationships formed between use and semantic traits—our recent visualization experiment underscores the vast array of meanings and the profound complexity of this concept.

The term "virtual" encompasses a full spectrum of meanings generated through use, some more clearly defined than others, some strongly and some loosely related. Drawing inspiration from Ludwig Wittgenstein's statements in paragraphs 66–69 of *Philosophical Investigations*, our analysis reveals that these semantic facets do not converge around a common core but instead form a complex network of relationships, described as family resemblances. We consider the construction of a panoramic view of the family resemblances for *virtual(ity)* a unique approach that serves to make a distinctive contribution to the field of Digital Humanities, enhancing our collective understanding of "virtuality" and expanding the existing body of scholarly knowledge. Our research methodology involved an extensive search process utilizing various tools, including integrations with Google Search API and scraping Google Scholar with third-party software. This approach yielded a substantial dataset comprising over 2000 results pertaining to the term "virtual." We subsequently employed this dataset to map and visually represent the intricate network of family resemblances, utilizing the Kumu platform.

Ludwig Wittgenstein's theory of "family resemblances," as elaborated in *Philosophical Investigations* (see also Ryan and Thon), offers a complex approach to

the human beings' conceptual understanding. The philosopher exemplifies his theory by looking at "game" and shows that when we investigate the different uses of the word game, we discover that while many correspondences can be found between uses, there are many features that drop out and others that appear. When looking at *board-games*, *card-games*, *chess*, *noughts and crosses*, *ball-games*, *a child's game of throwing a ball*, we should investigate *similarities* and *relations* in terms of *competition*, *amusement*, *winning and losing*, different types of *skill*, such as *skill in chess* and *skill in tennis*, *luck* etc. The similarities, correspondences and relationships between the features that appear or, in other cases, drop out form *a complicated network of similarities overlapping and criss-crossing: sometimes overall similarities, sometimes similarities of detail*. Built on this complex web of relationships, the term *family* is used by Wittgenstein to underscore his rejection of the idea that all members of a category should share a common core of similarities. In a family, for any given feature, there is at least one other member that shares it, but no trait is shared by all.[1]

Adopting a similar approach, we have delineated the network for "virtual(ity)," illustrating that each application of the concept unveils a distinct array of semantic nuances. These nuances then form a pattern of similarities or contrasts with other distinct applications. We note that certain features consistently emerge across uses while others recede. In addition to mapping similarities, we have introduced a contrasting relationship, particularly in relation to the concept *real*. The size of each node corresponds to the number of connections it has within the network, highlighting uses with extensive connections, such as *virtuality in fiction*, *virtuality of the real world*, *virtuality in storytelling*, and *virtualization*, as well as highly connected features like *real*, *digital*, *world*, *not actual*, *technology*, *possible worlds*, and *choice*. Through this method of visualization and prioritizing by connectivity, we show that

[1] The Family Resemblance Network for "game" following L. Wittgenstein and its analysis can be found here Stoenescu, Stoenescu, M. A. Family Resemblances: "game" Network Analysis. Zenodo Repository, 10 Dec. 2023, https://doi.org/10.5281/zeno-do.10340085.

nodes with high connectivity[2] represent uses that are more common and features that are generally considered more probable. Conversely, nodes with fewer connections can be deduced to be less frequent in usage. We can also observe the levels of relationships within this family by isolating relations for each node. We look at *virtuality in fiction* and identify on the map semantic traits such as *intentio lectoris*, *variation*, *time*, *choice*, *truth*, as well as the contrasting relation to *real*. Next, we see that each semantic feature is shared by other *uses* such as *text virtualization*, *VR, virtually possible.*[3]

Similar to the "game" analysis, our diagram illustrates the "family" concept which encapsulates the rejection of a universal semantic trait. For example, the term "virtual image" encompasses concepts such as *non-actuality* (A); *an image that appears to exist but does not* (B); *a specific phenomenon in optics* where an image is formed behind a mirror (C); *what is perceived visually* (D); and presents a contrast to what is *real* (E) and to *the actual state of affairs* (F), thereby constituting a feature set of {A, B, C, D, E, F}. In contrast, another usage such as *we are virtually there* conveys notions of *nearness to completion* (G), a *figurative sense* (H), *non-actuality* (A), *proximity* (I), resulting in the features {G, H, A, I}. Meanwhile, *virtual bank* implies a *digital* aspect (J) and involves *technology* (K), while still maintaining a contrast with the *real* (E), leading to the set {J, K, E}. The first and second sets share the concept of *non-actuality* (A), while the second and third sets share a contrast with the *real* (E). Although no single semantic trait is common across all three sets, we can detect one or more common features in some pairs. A key feature of the map is that it is dynamic and interactive; it allows for responsive exploration as hovering over a node highlights its immediate connections and dims the rest, thus facilitating an intuitive and informative navigation of the network.

[2] This frequency of use also influences the functioning of translation algorithms in machine translation tools.

[3] Snapshots of the different levels of the digram are available here: Stoenescu, M. A. Family Resemblance Network: Virtual - Levels 1 to 4. Zenodo, 14 Apr. 2024, https://doi.org/10.5281/zenodo.10970004.

As noted earlier, the dataset used to construct the Kumu map is derived from data obtained through meticulous examination and critical analysis of scientific literature on the subject. Specifically, the semantic traits (highlighted in blue) were manually gathered based on findings and descriptions from the research. While this method can produce a detailed map of the concept using selectively curated data, there's an argument to be made that such data may not genuinely reflect the authentic usage of the word. Consequently, we attempted to source a more objective dataset, aiming to capture a more organic linguistic application and a genuine representation of the concept. We believe that by examining the uses found in the results of the most dominant search engine, we can contend that acquiring a broader dataset in this way more accurately reflects the authentic application of the concept.

Tools for data retrieval (scraping), preprocessing, data analysis, and data visualization

Google API Integration: The optimization of search functionalities has become a pivotal concern in the contemporary digital landscape for researchers, scholars, developers, and engineers alike. As such, integrations with Google Cloud resources present an exceptional opportunity for scientific exploration and developing precise tools for scientific inquiry. Leveraging Google Cloud as an Infrastructure as a Service (IaaS) provides convenience and enhances precision, scalability, serverless computing capabilities, reliability, and access to comprehensive documentation for scientific projects. Without such a platform, scholars and developers would be constrained to construct and manage the required infrastructure themselves, a task that cannot match the services leading cloud providers offer.

The Google Cloud platform is accessible through the Google Cloud Console, where users can opt for more than 200 services and integrations covering a wide range of categories, including AI and ML (computing or conversational AI), containers, data analytics, databases, developer tools, networking, storage, and API management. For this project, we used the Google Search API to build a custom tool that uses the

Google search engine capabilities to retrieve results for specific queries. API stands for Application Programming Interface and consists of a set of protocols defining how different software applications can communicate with each other, e.g., data exchange, formats, and security. Generally, APIs provide a level of abstraction, hiding the underlying building blocks of a system. For that reason, they play a crucial role in software development by enabling developers and scholars to leverage existing services and functionalities to build new applications without programming everything from scratch. In our case, we used the Google Search API to integrate Google Search as a service in our script and tailor it to our requirements.

Setting up the access to the Google Search API requires following these steps. First, we create a new project in the Google Cloud Console. In the APIs & Services menu, we enable APIs & Services. We search for Custom Search API and enable it. Accessing the API through a command shell (or any script) requires a method of authentication, such as a personal API key. For that, we choose Credentials, Create Credentials, and API key from the menu. This unique key establishes the authenticated access to the API, ensuring a secure and authorized interaction with the service. For that reason, keeping it safe and secret is crucial. We stored it in a .txt file inside the project folder for later access.

With this, we have established the access to the Google Search service. The next step requires creating a Custom Search Engine (CSE)[4] that would perform the query task. For that, we follow the simple setup on programmablesearchengine.google.com. When creating a new CSE, a search engine ID is assigned. Like the API key, this ID is unique and should be kept safe and secret. We stored it in a new .txt file inside the project folder.

Building a Custom Scraper for search queries: When developing a software application, we typically rely on Jupyter Notebook. This widely recognized web application allows developers to create and share live code, equations, visualizations,

[4] "Custom Search JSON API." *Google Developers*, developers.google.com/custom-search/v1/introduction. Accessed 28 Sept. 2023.

and text. What sets it apart from other computing environments is its utilization of *cells*: distinct blocks of data that can be compiled independently. For example, developers can isolate a print function inside a separate cell, allowing them to fine-tune the desired output without having to recompile the entire script with every test. For that reason, it is particularly well-suited for data science and scientific computing projects, as well as for a wide range of Python projects. Its cell-based structure is not the only benefit of the Jupyter environment. Jupyter Notebook encourages users to include text explanations and documentation in their scripts. This is especially useful for data analysis and reporting, as well as for writing academic articles that include scripts.

For this project, we found the notebook's cell structure especially valuable due to the Google Search API limitation, which imposes a daily limit of 100 queries for the free plan. During the testing phase, without the ability to organize the query task in a separate cell, we would have quickly depleted our daily query limit each time we ran the script to test other functions. By isolating the query function, we ensured it was only executed when necessary.

Building a scraper for search queries Google Search API (libraries, key components, logic): The script (see Script 1 in Script Repository) extracts expressions containing "virtual," "virtually," "virtuality," or "virtualization" using the Google CSE through Google Search API[5], and saves the results, along with their context and links, into an Excel file for future process. The following is a detailed run through the script that justifies the choices and provides additional clarification. The main components are:

- API and CSE access
- search query setup
- search query task
- defining the regex pattern to extract results

[5] "Google Search API Overview." *Google Developers*, developers.google.com/custom-search/v1/overview. Accessed 25 Sept. 2023.

- listing the results
- output as .xlsx

We referenced three Python libraries: *requests* for the query function, *re* for the regex pattern, and *pandas* for data/file manipulation and storage. Next, we read the API key and the CSE ID from external files (API_KEY.txt and WittgensteinEngine_ID.txt) and stored them in two variables. This makes the approach more flexible (other search engines can be used by changing the ID) and more secure since sensitive data is not hardcoded into the script. Next, we set up the search query and the address to send the request. We performed many tests to optimize the search query considering the Google Advanced Search options (e.g., using operators like OR) as well as the limitations of the API, such as *pagination*[6] and the *number of results per page.* An example is the following search query: search_query = "virtual OR virtualization OR virtually OR virtuality." The script would send a query for the four words and generate a maximum of ten pages, with ten results per page, counting a total of 100 results per called function. This means that when using the OR operator, the 100 results are split for each of the four words. If we instead send a separate query for each of the four words, we get 100 results for each. For that reason, we chose to input the value of the query variable from the keyboard to get a more fine-tuned query setup. It could have also been done by using a *for* function and calling the query function for each of the words. Next, the query parameters are required for any query function.

[6] A note on pagination: Custom Google Engines return paginated results, which can be limiting when extracting larger datasets, not just the first results. To handle this limitation, the script sends multiple requests to the engine and aggregates results for each page. Unfortunately, this generates another issue regarding the consumption of the 100 daily queries credit that the free plan provides because when 'num' is set at 10 (pages), the script sends ten queries for each call. To better control the consumption for each run of the entire script, we chose to input the number of pages from the keyboard so that for simple tests, we would choose only one page, and for final runs, we would set it to the maximum value (10).

```
params = {
'q': search_query,          # The search_query variable[7]
'key': API_KEY, # Our private API_KEY
'cx': SEARCH_ENGINE_ID,
'start': 1,              # Start with the first result
'num': 10  # Number of results per page (max 10)
}
```

Regex patterns for raw data cleaning and preprocessing (challenges in the preprocessing phase): Next is the definition of the *regex pattern* we used to extract the desired expressions from the results. *Regex* stands for *regular expression*, a tool for pattern matching and text manipulation. Regular expressions are formed of a sequence of characters that define a search pattern that can be used for different operations such as searching, matching, extracting, validation, and replacement. The regex pattern we used is complex, but with the proper documentation[8] and enough testing, this is the pattern we finally created:

expression_pattern = re.compile(r'\b(\w+\s+virtual(?:ly|ity|ization)?)\s+(\w+(?:\s+\w+)?)\b', re.IGNORECASE)

This pattern captures expressions of the form: virtual + noun, virtually+word, virtuality+noun virtuality+in+noun, noun+virtualization, word+virtualization+noun, adj+virtualization, virtualization+of+noun. This regular expression is designed to match expressions of the form "virtual [word1] [word2]," where [word1] can be followed by optional suffixes "-ly," "-ity," or "-ization," and [word2] is another word. The capturing groups are used to extract [word1] and [word2] separately when a match is found.

[7] Here, the inline comments (marked with #) clarify the meaning of each variable. Moreover, the entire script includes comments that explain the purpose and functionality of different parts of the code, making it easier for others to understand and maintain the code.

[8] Tools such as https://regexr.com/, https://regex101.com/.

Octoparse integration for scraping Google Scholar (deployment and data retrieval): We used third-party software like Octoparse for our scraping tasks for two primary reasons. Firstly, third-party software solutions are often more advanced and robust compared to basic API calls for data extraction. In many scenarios, the challenge is not the API integration but processing the acquired data. Using third-party scrapers can often yield a richer dataset, which introduces intricate data-processing challenges worth exploring in a scientific paper. Secondly, upon examination, we found that Octoparse effectively addresses and overcomes certain limitations of the Google Search API and the Custom Search Engine, especially the 10-page limit and the restriction of 10 results per page. Interestingly, the Custom Search Engine interface mirrors Google's webpage design from a decade ago, featuring a top search bar, two tabs (web and image), an advertisement section, ten results, and a *pages* toolbar. In contrast, the current Google webpage has evolved significantly, now segmented into various sections such as dictionary, related search, "people also ask" section, and image section. Most notably, the traditional page toolbar has replaced the "More results" button. This evolution asks for a change in how scrapers interact with the webpage. We will look at Octoparse and highlight its pivotal features that enabled us to optimize data collection and bypass the restrictions associated with the API.

What follows is the documentation of the Octoparse experiment for scraping Google Scholar. There is an additional reason for introducing Octoparse here, which is the fact that there is no official Google Scholar API.

Octoparse is a web scraping tool that allows users to extract information from websites without writing code and building applications (scapers) from scratch. It is mainly designed for non-programmers and offers a visual interface where users can design their scraping tasks using predefined templates and setting up custom workflows. It is primarily used for data extraction of any kind (research, business intelligence, market analysis, news and content aggregation, etc.). Using a third-party tool like Octoparse helped us mitigate some of the limitations we have encountered with the Google API scraper. The advantages are many: no coding required, it has a visual workflow (users can see the data they are scraping in real-time so they can

fine-tune the tasks), no limitations for paginations or queries, the option to export raw data in xlsx, etc.

We also considered potential roadblocks and limitations when working with Octoparse (handling CAPTCHAs, limits, exports). Octoparse can be used in one of two ways: selecting a template or designing a custom task or workflow. The software does not allow you to download the data directly to your computer when using a template. Instead, it only offers the option to store it in its cloud storage, a feature that comes with a paid subscription. On the other hand, setting up a custom workflow is straightforward. It is similar to crafting a macro in MS Office. Essentially, you need to sequence the tasks a user would carry out during a Google search: launching the Google site in the browser, entering "virtual" in the search bar, hitting the *search* button or pressing enter, scrolling down, selecting the *More results* button, and so forth. Once established, this workflow acts as an "algorithm" that Octoparse runs automatically, capturing all the search results.

One particularly important aspect is that Octoparse does not use APIs, which raises the problem of handling CAPTCHAs. The reason for that is that many websites have security mechanisms that identify bots or invasive third-party software that access their data, such as web scrapers. CAPTCHA stands for "Completely Automated Public Turing Test to Tell Computers and Humans Apart." It is a system designed to determine whether the user is a human or a robot, and it is a prevalent method many websites use to combat scraping. While Octoparse does not handle it directly, there are a few ways to overcome this challenge.[9] The only method that proved effective for our project was using the *Browser mode*. This allows real-time observation of the scraper's interaction with the website, providing the option to pause the workflow and introduce user input directly into the site. Google is highly sensitive to third-party access and can swiftly detect when a scraper attempts to connect. Upon detection, Google triggers the reCAPTCHA system, often starting with the checkbox "I'm not a robot." Frequently, it follows this with another

[9] "How to Solve CAPTCHA in Octoparse." *Octoparse Help Center*, Octoparse, 2023, https://helpcenter.octoparse.com/en/articles/6471138-how-to-solve-captcha-in-octo-parse. Accessed 30 Sept. 2023.

CAPTCHA challenge: the Image CAPTCHA test. In this test, users are typically shown a prompt to select all images in a grid corresponding to a specific description, such as "Select all images with traffic lights." By temporarily pausing the scraper and addressing the CAPTCHA challenges, we successfully navigated past Google's protective measures.

After data capture, we implemented a series of filtering processes, applying the same regex pattern to both the title (the search result) and the snippet, using a subsequent Python script (refer to Script 2 below). Given its straightforward nature and similarity to the initial script, we will not focus on its specifics but will move to the project's next phase.

Matching *semantic features* with a *spaCy* model

The project's next phase involves linking the processed values to their corresponding semantic features. This step is likely the most challenging due to the size of the dataset and the complexity of matching uses to features. We have used various tools and experimented with several methods to automate this process.

ChatGPT 4 and Advanced Data Analysis Feature (Beta): GPT-4, the September 25 Version, incorporates four advanced tools that extend its functionality. The *Browse with Bing* feature enables real-time web searches for information retrieval; the *Advanced Data Analysis* feature enables complex data manipulation and analysis through Python code generation and execution; the *Plugins* feature enables different third-party plugins to enhance GPT-4, and the DALL·E tool, to generate images from textual descriptions (prompts) using a version of OpenAI's DALL·E image synthesis model.

The Advanced Data Analysis tool is one of the initial methods we have explored. The aim was to leverage the NLP and LLM capabilities of GPT-4 to analyze the set of semantic features and match them with the uses. This tool allows users to upload datasets (such as *xlsx* or *csv* formats) and specify the operations they want the AI to execute, or even describe the end result they aim for and let the AI handle the process. GPT-4 generates Python code in the background, which users can observe

in real-time. It executes this code within its own interpreter, debugging and modifying as needed to achieve the desired result. Once satisfied, users can instruct GPT-4 to save the refined dataset to a new file, which they can then download. The capability to upload, process, and download datasets circumvents the token limitations inherent in the prompting system, which is capped at 4,096 tokens for GPT-3.5 and 8,192 tokens for GPT-4. While these limits might appear very generous (approximately 32,000 characters per context window), processing large datasets directly within prompts leads GPT-4 to conserve space and computational resources. Consequently, it does not read or output entire datasets in full. To address the challenges posed by complex data processing and analysis tasks, the Advanced Data Analysis tool serves as an essential workaround. Nonetheless, this tool has a notable limitation. It performs the required operations by generating Python code and executing it, thus utilizing the LLM's capabilities solely for code writing (including debugging and optimization) but not within the code itself. For instance, if GPT-4 (in its default mode) is prompted to generate 20 semantic features for the word "virtual," it will use its LLM capabilities to do so effectively. However, when employing the Advanced Data Analysis tool for the same task, it will start generating the Python code and then execute it for this purpose. However, this code will rely on basic methods, such as generating sets of synonyms using basic models and libraries. If asked to perform a more complex linguistic analysis, it will commit to doing its best, yet it will still depend on basic modules and libraries. Admittedly, this tool can perform machine-learning tasks, which are inherently complex. But for our project, which aims to process a dataset that could later serve as a training model for a machine-learning algorithm, the lack of a pre-trained model to enhance our dataset means this is not a viable option. Having established that, we had to choose the best approach available, even if it meant leaving behind GPT's LLM capabilities, which would have been a powerful solution for our task.

WordNet similarity test: Our second approach involved writing an algorithm using the NLTK module, focusing on similarity calculations through WordNet's synsets, which are sets of synonyms sharing a common meaning. For each "use" of

"virtual" in the dataset, it calculates and compares its similarity to the *synset* of the semantic features (we predefined the semantic features set), recording the highest similarity score for each use-feature pair in the results. However, the script relies on WordNet's *synset* match for each word, which may not always represent the most contextually appropriate meaning, leading to potential inaccuracies in similarity scores. The reason is that semantic features are not synonyms but often incredibly nuanced meanings, specifically determined through use in a particular context. These sensible nuances may not even have a clear lexical expression, so trying to capture the semantic features with *synsets* is not the best approach. We used a similarity threshold of 0.25.[10] The total execution time was is around 15 minutes for our dataset (2000 rows and 150 columns approx.).

Similarity test with SpaCy word vectors: Shifting to spaCy, we utilized its word vector capabilities for semantic similarity computations. SpaCy's reliance on word vectors, which are multidimensional representations of words, allowed for capturing a broader array of semantic nuances. This approach provided a more nuanced understanding of context and word relationships than WordNet's path similarity. SpaCy also offered performance optimizations and faster text processing capabilities. Advanced features like part-of-speech tagging and entity recognition helped refine the process. We optimized the process by filtering out semantically less significant parts of speech, which reduced the computation time from about 90 seconds per iteration to 2-5 seconds. SpaCy's flexibility in optimizing functions for specific purposes proved beneficial in managing computation time. We used a strategy similar to the NLTK approach for measuring similarities, establishing a threshold to translate results into a binary format. The uniform distribution of values obtained with spaCy suggested a more precise and consistent evaluation of semantic relationships. After thorough testing, we set the optimal threshold at 0.5. Generally, spaCy is considered more accurate than WordNet when compared on summarization, and lexical richness calculation (Sharma, Aggarwal, and Alawadhi) and (Spring and Johnson). However,

[10] A higher threshold would result in fewer, more confident associations, while a lower threshold would include more associations but with less confidence.

in the case of semantical features as described by Wittgenstein, we note that "out-of-the-box" NLP libraries are still lower in accuracy compared to deep learning models using transformer architecture such as BERT or GPT.

Data visualization: final phase

Kumu[11] was our initial choice due to its user-friendly interface. Despite its advantages, like easy project sharing and customization, Kumu struggled with our project's scale, crashing even with reduced data (60% of 2000 nodes and 20,000 connections), compromising our goal to illustrate complex relationships in our study. Tableau,[12] known for its interactive dashboards and data analysis, was tested next. However, its limitation in plotting many-to-many node connections, crucial for network graphs, led us to rule it out. Finally, Gephi[13] proved to be the optimal choice for our project. As a leading open-source tool for graph and network visualization, it supports various file formats and offers powerful network analysis tools. Its high-performance rendering engine efficiently handles large networks, aligning with our needs for up to 50,000 nodes and 1,000,000 edges.

During the import step, Gephi requires (as any platform) a specific way of structuring the dataset: two dataframes — one for nodes and one for edges. Each node must have a unique ID, with optional tags for additional classification. We preprocessed our data, assigning an ID for each node and a type-tag (Use and Feature) to filter and format the uses and semantic features in the graph. The edges dataframe must list the connections between nodes on separate lines, with a 'Source' and 'Target' column containing the corresponding node IDs. Given that our plot comprises approximately 20,000 many-to-many connections, we had to represent

[11] "Documentation." *Kumu Help*, help.kumu.io/. Accessed 27 Sept. 2023.
[12] "Tableau Help." *Tableau Software*, help.tableau.com/current/pro/desktop/en-us/ default.htm. Accessed 25 Sept. 2023.
[13] "User Documentation." *Gephi - The Open Graph Viz Platform.* gephi.org/users/. Accessed 30 Sept. 2023.

each connection on a separate line in the edges dataframe. This resulted in multiple entries for individual nodes, totaling 19,586 lines.

After importing both dataframes into Gephi, we customized the graph for readability. Gephi is primarily built for network visualization and analysis; however, its default visualization features are not tailored to arranging nodes and edges in specific forms. Networks are typically intricate structures and often appear as free-floating forms, not regular shapes. Users can apply algorithms called *layouts*, such as Force Atlas, Fruchterman Reingold, or Yifan Hu to organize the elements or dictate their behavior. Visualizing the graph as two concentric circles (one for uses, one for features) seemed optimal to capture its essence. Since Gephi does not offer this type of layout by default, we integrated a plugin called Circular Layout, which provides three distinct layout algorithms: Circular Layout, Dual Circle Layout, and Radial Axis Layout. We primarily used the Dual Circle Layout to separate uses from features into two distinct concentric circles. Next, we customized the appearance of the nodes by filtering by the 'Type' tag (previously integrated into the dataset containing two values: use and feature) and selected two colors: orange for features and blue for uses. After enabling the visibility of the labels (the names of the nodes), we had to apply another set of algorithms because the nodes for uses were overlapping. We filtered out the features to display only the uses and applied the Expansion algorithm a few times, followed by the Noverlap algorithm and the Label Adjust algorithm. Finally, we plotted the graph at different resolutions and formats, yielding these results.

Figure 2 is the rendering of the entire network on a large scale.[14] The size of the feature nodes is proportional to their degree. The degree of a node is defined by its

[14] Due to this scale, individual nodes appear very small. Detailed images that offer a more zoomed-in view for better clarity are available here: Stoenescu, M. A. Total Family Resemblance Network: Virtual - 20000 Connections in Gephi. Zenodo, 14 Apr. 2024, https://doi.org/10.5281/zenodo.10970150. These perspectives enhance the readability of specific sections of the network and add depth to the representation, thereby illustrating the complexity and intricacies of the relationships within it. In particular, The view of the center closely resembles a weave or fabric due to the sheer volume of connections, creating an effect where the edges themselves seem indistinguishable, collectively forming a fabric-like structure. This representation invites multiple interpretations: it illustrates the dynamic and continually evolving nature of the concept, where new connections and features are perpetually integrated, further enriching the network; it signifies a high level of integration, suggesting that understanding one semantic feature likely leads to insights

number of connections, thus, nodes with a greater degree are represented as larger in size. This variation in size effectively embeds an additional dimension of information into the network's structure.

Conclusions

In our investigation into the concept of *virtual(ity)*, we have unearthed a complex network of semantic traits interconnected in a manner reminiscent of family resemblances. The nuanced variations of *virtual(ity)* across different contexts reveal a rich tapestry of meanings that go beyond a singular, definable essence.

Our study effectively harnesses the methodological framework of Wittgenstein's family resemblances concept to elucidate the semantic traits network surrounding *virtual(ity)*. We have confirmed Wittgenstein's argument that there is no universal common semantic trait across the entire set of meanings; however, many features such as *real*, *digital*, and *possible worlds* appear with notable frequency. The vast array of features underscores the multifaceted nature of the concept. Through a meticulous combination of scientific literature and organic data, we have constructed a more authentic linguistic representation of *virtual(ity)*. Integrating Google API, custom scraping techniques, and advanced data analysis using word vectors have been pivotal in this endeavor.

The research also thoroughly explored the application of various data visualization tools, with Gephi emerging as the preferred choice due to its robust handling of large datasets and complex many-to-many relationships. This visualization provides an aesthetic representation of the data and serves as an analytical tool to decipher the intricate semantic connections.

The broader implications of this research significantly extend beyond the semantic exploration of *virtual(ity)*. The methodologies and tools developed can be applied to other semantic studies, offering profound insights into the construction of

into others; and it implies that a segmented approach might not suffice for fully comprehending and analyzing this concept. Instead, a holistic view, considering the synergy and cumulative impact of various features, might be more effective in grasping the concept.

meaning in our increasingly digital world. Additionally, this research substantially contributes to computational linguistics and machine learning, providing a framework for semantic network analysis and the development of advanced translation algorithms.

Several open questions remain for future research. The dynamic nature of language and the evolution of *virtual(ity)* suggest that this network will continue to expand and shift. Further investigation into the impact of these shifts on translation tools and the exploration of additional semantic features as they emerge will be necessary. Moreover, the potential application of these findings in the development of artificial intelligence and machine learning models holds immense promise for future inquiry.

In conclusion, our comprehensive journey through the semantic landscape of *virtual(ity)* has demonstrated a complex web of meanings, decisively challenging the notion of a singular definition. The application of the *family resemblances* framework has not only been exceptionally effective in mapping this intricate terrain but also has revealed the dynamic and continually evolving nature of the concept advocating for a holistic view, one that embraces the synergy and cumulative impact of various features, to effectively grasp the full concept.

Moreover, the family resemblances framework has not only mapped the existing semantic landscape but also provided a robust methodological foundation for future explorations in computational linguistics and machine learning. The insights garnered from this study have major implications for these fields, offering a new lens through which the construction and evolution of complex concepts can be understood and applied.

As we progress in our research, we are committed to continuously refining our methodologies and analytical tools to represent the complexity of *virtual(ity)* more accurately. This research highlights the potential for significant advancements in understanding digital phenomena. The future presents numerous opportunities for further scholarly inquiry and methodological innovation in this field.

Accordingly, we posit that the exploration of *virtual(ity)* extends beyond mere academic inquiry; it is an investigation into a concept central to our increasingly digitalized world, with substantial implications for comprehending the foundational

concepts that shape our contemporary experience. The depth of our findings encourages further examination, and we invite the academic community to join us in this pursuit, contributing to a richer, more nuanced understanding of the virtual and its myriad dimensions.

Annex 1: Figures

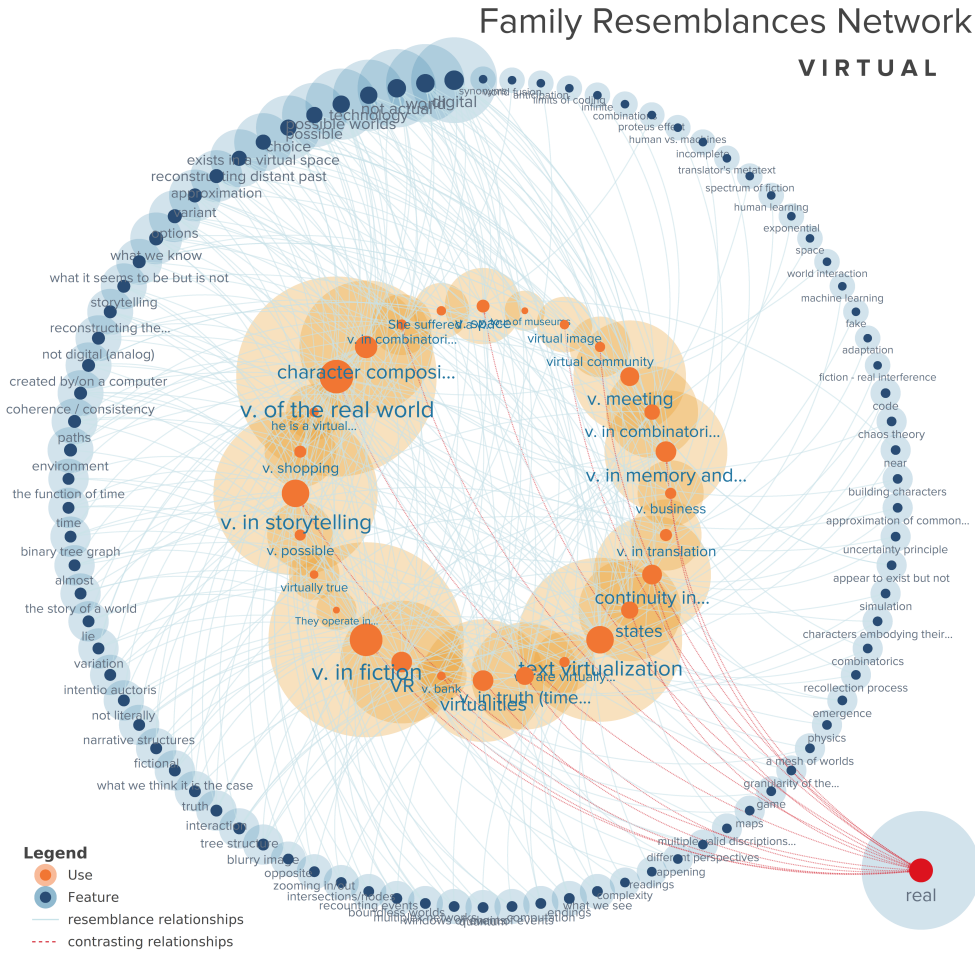


Figure 1: Family Resemblance Network: virtual.
The Dynamic model is available at https://embed.kumu.io/63c4468fe768675f5fc1960ffb6b87cf.

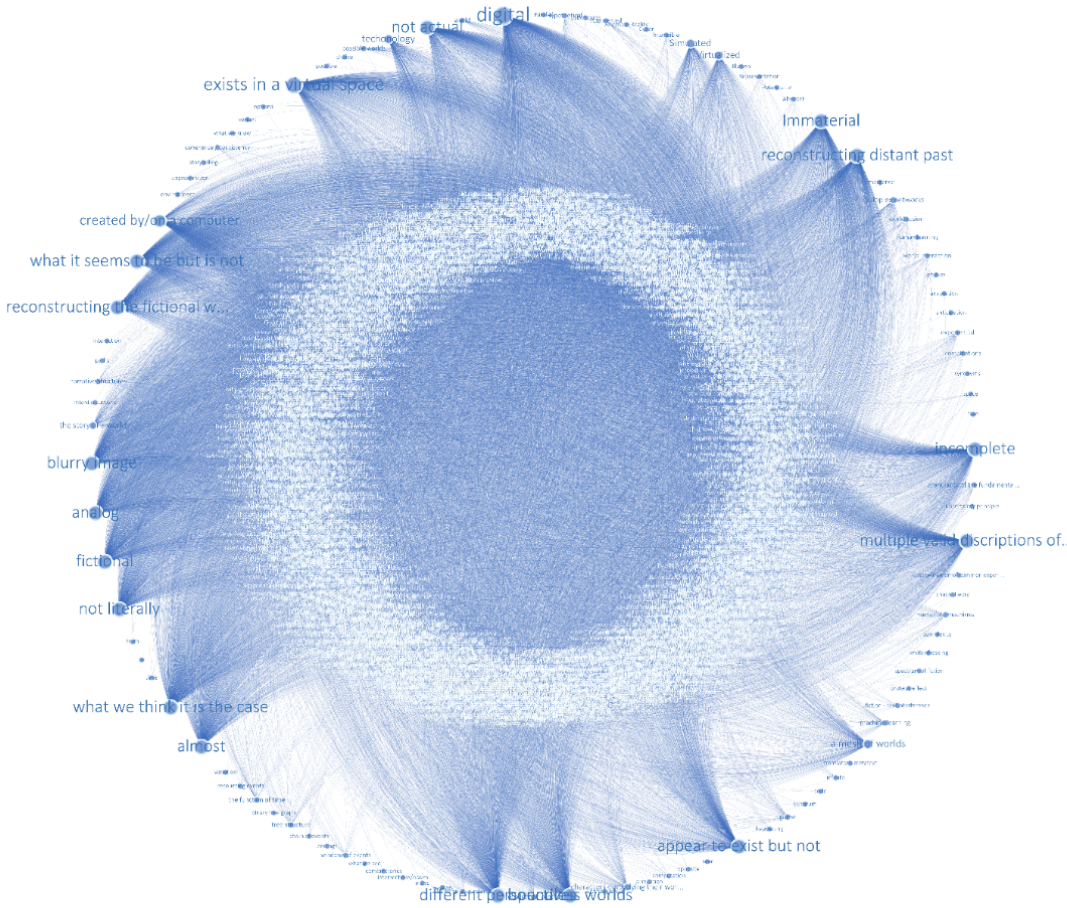


Figure 2: Total Family Resemblance Network – virtual – Gephi

Annex 2: Script Repository

Script 1

```
#This is the main script.
#It uses Google Search API integration through Custom Search Engine to retrieve data from Google Search.
#It writes one time (then overwrites) to an excel file the Expression (regex extracted) + Context + links
#for query = 'virtual OR virtually OR virtuality OR virtualization'

import requests
import re
import pandas as pd

#Read from file the API key and the Search Engine ID. Both are private keys and will not be disclose here.
API_KEY = open('API_KEY.txt').read()
SEARCH_ENGINE_ID = open('WittgensteinEngine_ID.txt').read()

#Set up a search query and the url to send the request to (our custom Google search engine)
# Take user input for the search query
search_query = input("Enter your search query: ")
base_url = "https://www.googleapis.com/customsearch/v1"

#Set up the parameters of the request as a dictionary of parameters' key' as  and 'cx' is the ID of our Search Engine
params = {
```

```
    'q': search_query, # The search_query variable
    'key': API_KEY, # Our private API_KEY
    'cx': SEARCH_ENGINE_ID,
    'start': 1, # Start with the first result
    'num': 10 # Number of results per page (max 10)
}

# Initialize lists to store extracted expressions and their corresponding links
all_results = []
extracted_expressions = []
extracted_links = []

# Take user input for the number of pages to retrieve
num_pages = int(input("Enter the number of pages to retrieve: "))

# Send the requests to the url with the parameters that we defined for multiple pages
for page in range(num_pages):
    response = requests.get(base_url, params=params)
    results = response.json().get('items', [])
    # Add results from this page to the list of all results
    all_results.extend(results)

    # Increment the 'start' parameter to get the next page of results
    params['start'] += 10

# Define a regex pattern to extract expressions of the form "virtual [noun]"
expression_pattern =
re.compile(r'\b(\w+\s+virtual(?:ly|ity|ization)?)\s+(\w+(?:\s+\w+)?)\b', re.IGNORECASE)

# Initialize lists to store extracted expressions, their corresponding contexts
(snippets), and links
extracted_expressions = []
extracted_contexts = []
```

```
    extracted_links = []

    # Extract expressions from titles, snippets, and links
    for item in all_results:
        title = item['title']
        snippet = item.get('snippet', '')
        link = item.get('link', '')

        # Search for the expressions in the title and snippet
        title_matches = expression_pattern.findall(title)
        snippet_matches = expression_pattern.findall(snippet)

        # Add the matches to the list of extracted expressions, contexts, and their
corresponding links
        for expression in title_matches:
            extracted_expressions.append(' '.join(expression))
            extracted_contexts.append(snippet)
            extracted_links.append(link)
        for expression in snippet_matches:
            extracted_expressions.append(' '.join(expression))
            extracted_contexts.append(snippet)
            extracted_links.append(link)

    # Create a DataFrame with columns for expressions, contexts, and links
    df = pd.DataFrame({'Expressions': extracted_expressions, 'Context':
extracted_contexts, 'Links': extracted_links})

    # Save the DataFrame to an Excel file
    df.to_excel('virtualization_extraction1.xlsx', index=False)
```

Script 2

#This Script reads the excel file containing the raw data extracted from Google Scholar, extracts with Regex the desired data from the Title and the Snippet and stores them in a new excel file.

```
import pandas as pd
import re

# Load the original Excel file into a DataFrame
original_file_path = 'GoogleScholarExtraction-raw_work.xlsx'
original_df = pd.read_excel(original_file_path)

# Extract the first and second columns into separate lists
column1_values = original_df.iloc[:, 0].tolist()
column2_values = original_df.iloc[:, 1].tolist()

# Initialize lists to store the extracted data and snippets
extracted_data = []
snippets = []

# Define the regex pattern
pattern = r'\b(\w+\s+virtual(?:ly|ity|ization)?)\s+(\w+(?:\s+\w+)?)\b'

# Iterate through the lists and extract data using regex
for value1, value2, snippet in zip(column1_values, column2_values,
original_df['snippet']):
    match1 = re.search(pattern, str(value1))
    match2 = re.search(pattern, str(value2))

    if match1:
        extracted_data.append(match1.group(0))
        snippets.append(snippet)
```

```
        if match2:
            extracted_data.append(match2.group(0))
            snippets.append(snippet)

    # Create a new DataFrame with the extracted data and snippets
    new_df = pd.DataFrame({'Extracted_Data': extracted_data, 'Snippet': snippets})

    # Save the new DataFrame to a new Excel file
    new_file_path = 'ExtractedDataWithSnippet.xlsx'
    new_df.to_excel(new_file_path, index=False)

    print(f"Extracted data with 'snippet' column added saved to '{new_file_path}'")
    Extracted data with 'snippet' column added saved to
'ExtractedDataWithSnippet.xlsx'
```

Script 3

```
    # This script uses spaCy module to match uses to semantic features.
    # Step 1: Setup
    import spacy
    import pandas as pd
    import numpy as np

    # Load the pre-trained spaCy model
    nlp = spacy.load("en_core_web_md")

    # Load the dataset
    df = pd.read_excel("ExtractedDataWithSnippet_work.xlsx")

    # Adjust the tokenize function to return SpaCy tokens instead of strings
    def tokenize(text):
```

```
        return nlp(text)

    # Compute similarity only if tokens have vectors and are not stop words
    def compute_similarity(token1, token2):
        if token1.has_vector and token2.has_vector and not (token1.is_stop or
token2.is_stop):
            return token1.similarity(token2)
        return 0

    # Function to process each row and calculate similarity, excluding prepositions and
conjunctions
    def process_row(row, features):
        row_data = []
        tokens = [token for token in tokenize(row['Label']) if token.pos_ not in ('ADP',
'CONJ', 'CCONJ')]
        for feature in features:
            feature_tokens = [token for token in tokenize(feature) if token.pos_ not in
('ADP', 'CONJ', 'CCONJ')]
            max_similarity = 0
            for token1 in tokens:
                for token2 in feature_tokens:
                    max_similarity = max(max_similarity, compute_similarity(token1,
token2))
            row_data.append(max_similarity)
        return row_data

    # Define the features to compute similarity
    features = df.columns[2:-1]  # Excluding the last column

    # Starting similarity computation
    print("Starting similarity computation for each use-feature pair...")
    results = []
    for index, row in df.iterrows():
        row_data = process_row(row, features)
```

```
            results.append(row_data)
            # Print progress
            print(f"Processed {index + 1} of {len(df)} rows...", end='\r')

        print("\nFinished similarity computation for the loaded rows.")

        # Step 4: Thresholding
        threshold = 0.5
        true_matrix = (np.array(results) > threshold).astype(int)

        # Create a new DataFrame from the true_matrix to avoid fragmentation
        true_df = pd.DataFrame(true_matrix, columns=features)

        # Concatenate the new DataFrame with the original one, this will create non-
fragmented DataFrame
        df = pd.concat([df.iloc[:, :2], true_df], axis=1)

        # Calculate the total associations and add as a new column
        df['Total_Associations'] = true_df.sum(axis=1)

        # Save/display results
        df.to_excel("updated_data_spacy_threshold_0_5_first_allows.xlsx", index=False)
        print(df)
```

References

Aarseth, Espen. 2001. "Virtual Worlds, Real Knowledge: Towards a Hermeneutics of Virtuality." *European Review* 9 (2): 227–232.

Bartle, Richard Allan. 2004. *Designing Virtual Worlds*. Indianapolis: New Riders Publishing.

Belanger, Wayne D. 2009. *A Semiotic Analysis of Virtual Reality*. UMI Microform, PhD dissertation.

Chan, Melanie. 2014. *Virtual Reality: Representations in Contemporary Media*. London-New York: Bloomsbury.

Heim, Michael. 1993. *The Metaphysics of Virtual Reality*. Oxford: Oxford University Press.

Heim, Michael. 1998. *Virtual Realism*. Oxford: Oxford University Press.

Lewis, David. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Ryan, Marie-Laure, and Jan-Noël Thon. 2014. *Storyworlds Across Media: Toward a Media-Conscious Narratology*. Lincoln-Nebraska: University of Nebraska Press.

Ryan, Marie-Laure, and Alice Bell (eds). 2019. "From Possible Worlds to Storyworlds: On the Worldness of Narrative Representation." In *Possible Worlds Theory and Contemporary Narratology*, edited by Alice Bell and Marie-Laure Ryan. Lincoln-Nebraska: University of Nebraska Press, 62–87. Accessed January 14, 2021. JSTOR, www.jstor.org/stable/j.ctv8xng0c.7.

Wittgenstein, Ludwig. 1953. *Philosophical Investigations*. Translated by G.E.M. Anscombe. New York-London: Macmillan.

Google Developers. 2023. "Custom Search JSON API." Accessed September 28. developers.google.com/custom-search/v1/introduction.

Google Developers. 2023. "Google Search API Overview." Accessed September 25. developers.google.com/custom-search/v1/overview.

Octoparse. 2023. "How to Solve CAPTCHA in Octoparse." *Octoparse Help Center*. Accessed November 1. helpcenter.octoparse.com/en/articles/6471138-how-to-solve-captcha-in-octoparse.

Sharma, Abhilasha, Raghav Aggarwal, and Raghav Alawadhi. 2023. "A Comparative Study of Text Summarization using Gensim, NLTK, Spacy, and Sumy Libraries." *Journal of Xi'an Shiyou University, Natural Science Edition* 19 (4).

spaCy Documentation. 2023. "spaCy 101: Everything you need to know." Accessed September 30. spacy.io/usage/spacy-101.

Spring, R., and M. Johnson. 2022. "The Possibility of Improving Automated Calculation of Measures of Lexical Richness for EFL Writing: A Comparison of the LCA, NLTK and SpaCy Tools." *System* 106. https://doi.org/10.1016/J.SYSTEM.2022.102770.

Tableau Software. 2023. "Tableau Help." Accessed September 25. help.tableau.com/current/pro/desktop/en-us/default.htm.

Gephi. 2023. “User Documentation.” *Gephi - The Open Graph Viz Platform*. Accessed September 30. gephi.org/users/.
Kumu Help. 2023. “Documentation.” *Kumu Help*. Accessed September 27. help.kumu.io/.

CARTOGRAFHIES OF THE ROMANIAN NOVEL: THEN AND NOW

The Recontextualization of a Genre[1]

Loredana CUZMICI

The general impression regarding nineteenth-century Romanian literature has been that shorter works, such as short fiction and poetry, dominated it. This was partially due to the famous debate started by Nicolae Iorga in 1890 and titled "*De ce nu avem roman?*" [Why Do We Not Have a Novel?],[2] in which economic reasons took center-stage (Iorga in Sasu and Vartic 1997, 24–28). The Romanian historian and the public personalities joining the debate claimed that novels were not written in this side of Europe because literature was practiced by non-professional writers as a leisurely activity. The discussion went on by considering the contents of the "great" histories of Romanian literature, or educational and academic curricula. The question persisted for more than three decades, being echoed by both similar and opposite responses. Beside the socio-economic precariousness of the newly-formed state, later explanations for the delayed development of the novel took into account imagological aspects—for instance, the Romanians' "transactional spirit" and non-conflictual temperament was perceived as a drawback for the novel's narrative climax—, the fragile cultural tradition, including the absence of a traditional epic model (Sasu and Vartic 1997, 74–81), the excessive melodrama, the lack of psychological complexity, or the subpar construction of the protagonists, the latter being considered by novelist and philosopher Camil Petrescu a fundamental element around which the plot coagulates (in Petrescu in Sasu and Vartic 1997, 82–83). Nevertheless, efforts to recover, index, and reproduce nineteenth-century publications for the general public gave cause for surprise: in reality, the novel dominates the Romanian context, as

[1] This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS – UEFISCDI, project number PN-III-P1-1.1-TE-2019-0127, within PNCDI III.

[2] All original quotes are translated by the author of this chapter.

much as it did the rest of Europe. Quantitatively, the majority of publications were novels, whether they were volumes *per se* or instalments/ series/ feuilletons of different journals.

It was Mihail Kogălniceanu's 1840 warning that "*Traducţiile nu fac o literatură*" [Translations do not make a literature], combined with Titu Maiorescu's famous esthetic criticism a few years later, that made it so that, obviously influenced by western models, popular literature from the nineteenth century and from the first two decades of the twentieth century was either nearly disregarded in studies of literary history, or presented with many caveats. Marked by national identity (which was boosted after the Union of the Principalities, the Independence, and the Great Union after World War I), the sociopolitical ideals of the most important cultural figures of the time exerted a great influence over all public manifestations. This trend resulted in the fact that popular literature, because of its cosmopolitan airs, gained nothing but poor critical reviews, despite the fact that the majority of books offered localized plots and protagonists.

When one investigates the important histories of Romanian literature, the main ideas concerning nineteenth century revolve around the influence of the two geo-cultural axes, East-West, with a scale that leans ever more towards the latter. G. Călinescu, whose history remains the fundamental reference for the evolution of Romanian literature up to World War II, starts his analysis of nineteenth-century literature with the suggestive title, "*Descoperirea Occidentului*" [The Discovery of the West] (Călinescu 1986, 61). Up until the interwar period, one of the trademarks of Romanian culture is its permanent benchmarking against the Western model, which becomes a criterion for value. However, a relatively ignored criterion, but one that has been resurfacing over the past few decades, is public reception. What did Romanian literature look like during the Century of Nations, if we consider it from the perspective of amateur readers, rather than from that of critics or writers? One thing is for certain—the public were not only reading the works belonging to Romanian and foreign authors that would later be canonized, but also writings that satisfied the general taste: short fiction, embellished with various illustrations, some initially published in instalments and tackling sensational themes, counting on

suspense and surprise. Intrigue and character construction followed certain patterns, while dominant themes, indexed with capital letters by writer and professor Andrei Bodiu, were Romance (corresponding to the somewhat undervalued alternative to romantic love), Society, the City, Suffering, Death, the Other, and, naturally, Writing (Bodiu 2002).

In the first half of the nineteenth century, Ion Heliade-Rădulescu launched his famous incentive, which boosted the publication of a variety of texts: "*Scrieți, băieți, numai scrieți!*" [Write, boys, just write]. Esthetic value was not a priority just yet, although the poet and publicist Heliade-Rădulescu is counted as one of the great writers of his time and of Romanian culture. The imperative was that empty spaces be filled quickly and rapid modernization be carried out through the so-called "burning stages", which would have an effect including on the production of popular mass-market fiction. This is a highly specific context: while in the rest of Europe, there was already a discussion about a "second modernism" occurring in the second half of the nineteenth century (Petrescu 1992), within the Romanian cultural context the two initial stages of the novel overlap, with the first taking precedence, and the second formula appearing almost accidentally, particularly in localized translations (Teliban 2024) such as *Catastihul amorului* and *La gura sobei* ["The Catastrophe of Love" and "By the Hearth"]. The advancement of the press, of educational institutions and the theater, alongside the more and more frequent travels and studies abroad, allowed for massive imports, which, illustrating the famous theory of cultural imitation (Lovinescu 1981, 319–326), gradually stimulated the local production of texts.

Prior to research conducted after World War II, "the battle for the novel" (Sasu and Vartic 1997) represents a significant moment in the analysis and evolution of Romanian literature, encompassing multiple decades, starting with late nineteenth century and enhanced during the interwar period. The predominant tenet was that the number of published novels was relatively low up until 1920, the year Liviu Rebreanu's famous novel *Ion* was first published (the same year thought to mark the end of *the long nineteenth century*, *see* Hobsbawm 1995 and Burke 2000). Also, the

esthetic quality of these early-stage novels was highly questionable. In truth, we did have a novel, but we also had several preconceptions regarding the concept of literature. Only a handful of novels were subsequently canonized, with the most representative in anthologies being *Manoil* and *Elena* by Dimitrie Bolintineanu, *Ciocoii vechi și noi* [Old and New Arrivists] by Nicolae Filimon, *Mara* by Ioan Slavici, *Viața la țară* [Life in the Countryside], and *Tănase Scatiu* by Duiliu Zamfirescu. Alternatively, countless others belonging to "the genre" of the *popular novel/ fiction* were disregarded by critics, themselves in the early stages of their craft. However, these marketable publications were preferred by the bourgeoisie, a new rising social class that was eager for identification and relatable reading. Many novels would thus be forgotten in the pages of various publications and are yet to be rediscovered by researchers. It is noteworthy that the very titles which would later be categorized as top-tier in terms of value had initially come out in serialized format, and only later in volumes. Checking success was a mechanism spread all around Europe: firstly, the novels would be "approved" by the general public, that is, the target audience of magazines and newspapers; secondly, the novels would be published in volumes. The nineteenth century is defined, first and foremost, by the enormous development of the press, which, by stimulating an interest in literacy, would further boost a taste for literature. Along with the growth of press as "the nineteenth-century media matrix" (Thérenty 2007), the novel manifests from the very beginning as a democratic genre. Additionally, "the age of mechanical reproduction" (Benjamin 2015) would also have immediate effects on spreading fictional worlds.

In search of a national tradition and specificity, the beginning of the twentieth century and the first interwar years would thus be marked by epistemological interrogation and cultural revisionism. While Romanians were pondering on the reason why they did not have a novel, the Russian Formalists got to the root of the concepts by asking what literature was and proposing one of the most interesting linguistic and poetic theories, with countless echoes in subsequent literary and cultural currents. The concept of *literariness* advanced by Jakobson as the principal trait of literary texts was no longer a prerogative reserved to masterpieces. The so-called *paraliterature* began to elicit the interest of Russian theorists, for whom the

esthetic criterion, fundamental in the literary criticism of the times, was no longer the only means to legitimize a text. In contrast with the peaks of value, whose expressive functions, both poetic as well as metalinguistic, represented a priority, mass-market novels are oriented by the phatic, conative, and referential functions (Drăgan 2006, 44). If we consider Constantin Dobrogeanu-Gherea's or G. Ibrăileanu's opinions, in the Romanian context there was also a position who promoted the heteronomy of literature and its release from the (classical) tyranny of the esthetic. However, the two socialist critics paid little attention to popular literature, categorizing it as far too tributary to foreign models. With a totally different grip, Nicolae Iorga and Mihail Sadoveanu mention the formative readings they did, particularly during childhood and adolescence, including novels with wide public reception and circulation (Pillat 1998, 128–144; Sadoveanu 1970, 120–126; Iorga 1936). After all, unlike popular books that had been in circulation even during prior centuries (*e.g. Alexandria*, *Sindipa*, *Floarea darurilor* [Flower of Virtue], apocryphal religious writings, etc.), the popular novels of the nineteenth century offered something more: their authors would, to a great extent, follow the will of the public, adapting their themes, bringing them closer to contemporary existence, so as to ensure a reading of identification. What is more, the religious perspective was supplanted by a secular one, and the pronounced didacticism, common to popular books of former centuries, would be diminished.

It was not until after World War II that critics such as Dinu Pillat (encouraged by G. Călinescu), Paul Cornea, Nicolae Manolescu, Marian Barbu, Cornelia Ștefănescu, Mircea Popa, and others would turn their attention to mass literature. Adrian Marino, in his theoretical syntheses, emphasizes the importance of studying *paraliterature* (Marino 1998, 119). In the context of theoretical developments regarding world literature and the reconsideration of canons, a keen interest in the unseen part of the literary iceberg has been growing in recent years. Marginal genres, such as the detective novel or children's literature, are now carefully studied, particularly since the appetence of the general public is largely shaped by such readings. Literature is analyzed especially through the lens of its social relevance, as

an "epistemological metaphor", to quote Umberto Eco, while esthetic criteria take second place (Eco 2006, 162–168).

A moment of utmost importance in the study of this phenomenon is the editing of *Dicționarul cronologic al romanului românesc* [The Chronological Dictionary of the Romanian Novel] (hereafter DCRR), coordinated by Ion Istrate, a volume published in 2004 which follows the evolution of the genre since its inception up to 1989[3]. We can, therefore, notice that it is only in the twenty first century that a major synthesis of the Romanian novel appears, a piece of longitudinal research conducted by a team of researchers from the Institute of Linguistics and Literary History "Sextil Pușcariu" in Cluj-Napoca. Likewise, following in the footsteps of their mentor Paul Cornea, researchers like Ioana Drăgan and Daniela Zeca study the way in which nineteenth-century literature and *paraliterature* influenced each other, resulting in a dynamic, varied landscape, and reflecting the spectacular economic and social changes undergone by two states (the Romanian Principalities of Moldova and Wallachia) that had recently emerged out of medievalism.

Starting from DCRR and based on its groundbreaking view on the evolution of the Romanian novel, digital archives and corpora of novelistic texts have been constructed within the past decade. They include enough titles so as to alter the way in which we used to look at the history of local novels and to define a critical position opposing the aforementioned perspective of early debaters on the status of the novel. The recovery effort extended over "the great unread" has become a world-wide endeavor over the past few years, with certifications such as Franco Moretti's "atlas of the European novel". Moretti's model relies on expanding the study of the "genre" from its history and biology acceptations, to the geography of the novel. In line with these developments, we notice canon revisions happening within Romanian literature as well, all of them meant to diversify the criteria for literary text analysis and to reevaluate forgotten writings.

[3] Meanwhile, a second edition revised and expanded has recently been published (see *Dicționarul cronologic al romanului românesc de la origini până în 2000*, vol. I–II, ediție revizuită și adăugită, Cluj-Napoca, Presa Universitară Clujeană, 2023).

A deleterious prejudice: the limits of the esthetic

Prior to the avant-garde that took poetry to the streets and proclaimed the emotional and social ends of art, the mass literature of the long nineteenth century was already accessible to a wide range of readers, particularly the ladies of the bourgeoisie, eager for fictional escape. A discreet and progressive revolution was stirred among the population, with novels decidedly contributing to the emancipation of certain social categories, much more so than poetry or short prose. Equally important were theatrical representations, whose effects were just spectacular, thus not catered for reading responses. The mainly emotional function of this type of literature and the emphasis on enjoyment were meant to part with the categorical imperatives of the classical esthetic; reading as a psychological act, not as cultural one, became a common practice among the literate population. While art for art's sake was of no interest outside elite groups, "art with a tendency" dominated, but it was precisely the coexistence of these two directions that offered dynamism and a sense of progression. As such, long before theorists would bid "farewell to aesthetics" (*see* Jean-Marie Schaeffer 2000) and before the fomenting of worldwide revolutions against the (Western, official) canon or even before the counterculture movements, *run-of-the-mill* novelists dared to stand against the lofty imperatives of expressive writing and to produce a variety of texts that would be relegated after a few decades only to library databases or mentioned only in *varia* chapters. Naturally, this phenomenon is widespread across all cultures, be they major or minor.

Why, then, would these forgotten archives be of interest to us nowadays? The answer has been formulated in recent contributions to the field of computational literary studies and can be summarized as follows: because the thickness of archives gauges a social barometer, as esthetic realization does not always equate with a failed act. Therefore, the ontological function of literature can be demonstrated even in this case of quantitative methods like distant reading, which cannot be resumed to instrumentalism or mechanical objectivity (Bode 2023).

The "bitterness of aestheticism" lamented by Camil Petrescu, who accused overly-refined literature of artificiality, of a "deleterious distancing from life and

reality", may therefore be tested even before the emergence of the imperative of authenticity. The novelists of the nineteenth century did not shy away from the seduction of fantasy (ascertained as one of the human being's most authentic needs) or from exaggeration and overemphasis. Building fictional worlds, albeit considered implausible from multiple points of view, was perceived as an enjoyable activity nonetheless. The disdain against *phantasia*, against the so-called escapist literature, far too conventional and simplistic, expressed by all followers of "High" Realism (by association with High Romanticism) would give a leeway to reservations or even skepticism regarding these popular novels. Popular success was automatically equated with non-value, and this radical, unilateral perception impoverished the landscape of Romanian literature. Despite their being listed at the top of all canons, it is not only novels of experience, profoundly autobiographical in their nature, that have a right to exist. This plea for the mass novel may be traced in one of the most commented novels of the interwar period, *Adela* by G. Ibrăileanu. The protagonists of the novel, two lovers, have a debate over the value of certain texts, their readings of choice becoming an effective method for cross-portraying instances. The female character wonders why great writers choose to complicate the existence of their protagonists, why an unhappy denouement would be preferable, why the idyllic and idealism are thought of as false. Ultimately, from Adela's discourse, we can extract a subtle criticism against realism (in fact, a demythologizing, deromanticizing formula), which limits imagination and obliterates illusion (Ibrăileanu 2010, 785–786).

What is more, there are no crises of language, unsurmountable ontological dramas, acute perceptions over the absurdity of the world or other elements that generally define high literature to be found in mass novels. The so-called superficiality of their subjects made them all the more accessible. Indeed, from its inception in the European context, the novel was mostly regarded as a plebeian form, if not downright vulgar, a synecdoche for subliterature. Despite quickly taking the very front lines of literature, it maintained consistent collateral values with the so-called *Trivialliteratur*. The convention of identification with the real—by the use of the well-known paratextual formula "a true history", the motif of the discovered

manuscript, characters drawing from real historical figures, toponyms, etc.—attracted more readership than the technical sophistication spread over hundreds of pages in "great" novels.

The impact literature has on society is much more visible if we take into account popular productions. As a rule, masterpieces (with all the reservations we have regarding these absolute terms) extend beyond their age, whereas serialized writings, endorsed (particularly through purchasing) by a wide audience are a true reflection of the mentality of their age and provide, therefore, a far more generous source for studying a cultural context. Since the rigors of exclusive esthetics are no longer the sole principles of literary analysis, the immense role exerted by mass literature on social emancipation justifies the attention it has garnered over the past decades. A further reason for the importance of studying this macrophenomenon is the propensity of mass literature to metamorphosize into propaganda literature, a phenomenon that occurs in the first decade of Romanian communism, when the socialist-realist formula was the only option for writers.

On a different note, it has been argued that mass literature adopts techniques and even themes from great literature, by simplifying them. However, in an apparently paradoxical turn, *paraliterature* assists high literature. First of all, it forms readers, who refine their taste as they explore various textual offers. Just as writers practice typing with their first texts, preparing for their "great" oeuvres, so do readers practice reading. After all, what would Romanian writers Caragiale or Urmuz have parodied had there not been an inflation of romance novels in their times? Writers such as these two deconstruct fashionable formulas, sanctioning clichés and conventions, which result in a literature-of-the-second-degree whose mechanism cannot be interpreted without the hypotexts that they spoof or parody. Therefore, as demonstrated by Ioana Drăgan, there is an interdependence of the two literary fields: the novel is a literary genre situated at the frontier between the two circuits, that is, *the literary* and *the paraliterary* (Drăgan 2006, 9). Even with all the derogatory names it was bestowed over time — "industrial literature", "*littérature de bas-étage*", "dime novels", "throw-away literature", "pulp fiction" (Drăgan 2016, 19 49)—not only is

this type of literature surviving, but it is expanding and diversifying its forms. Nowadays, public success means that many of these *dime* texts are adapted for the screens, and their impact on mentalities becomes even more visible. After all, these aspects demonstrate the inherently democratic spirit of the novel: "it is what ethnologists have confirmed by observing that novelistic literature is unknown in caste societies or primitive peoples whose social structures are immovably fixed in tradition" (Robert 1983, 68).

For the wealth of nineteenth-century texts (corpora of Romanian novels *y compris*), quantitative methods, energized by the omnibus concept *distant reading*, are the most effective study method, allowing for the extraction of metadata, observing invariants, as the enormous quantity of texts becomes nearly impossible to cover through classical qualitative methods.

Novel subgenres from the nineteenth century to 1920

In Romanian novel theory, among the most operational and well-known classifications are those belonging to G. Ibrăileanu (1930), N. Manolescu (1980), and Radu G. Țeposu (1983). They have the advantage of simplifying a terminology which might otherwise risk a tentacular spread, as well as espousing aspects of narrative technique rather than merely thematic or chronological ones as classification criteria. The "creative", the "Doric" or the "transitive" novel would be specific to the emerging phase of the genre, easily defined using the metaphorical formulas given by famous titles, such as "the human comedy" and "vanity fair", entirely unbeknownst to the characters. Instead, the author's vision is largely projected on the narrator, who might be either critical, or, less commonly, impersonal. Thus, society as a whole becomes the protagonist of this type of novel. Made up of representative figures, it often acts as an oppressive force on the individual, or in other cases as a savior, as meticulously demonstrated by Nicolae Manolescu in one of the most encompassing studies on the evolution of the novel. The omniscient narrator remains the chief narrative instance, whose point of view dominates (Manolescu 1998). The "creative" novel for which G. Ibrăileanu advocates in his study *Creație și analiză*

[Creation and Analysis] (Ibrăileanu 2010, 605–637), consubstantial to a great extent with the Doric novel, is represented by Tolstoy's formula in *War and Peace*: extensive scenes, multiple planes of narration, strong conflicts, without excluding analysis, but rather subordinating it to action. An organic parallel becomes unavoidable, as this first form of the novel aligns with a masculine representation of the world, in competition with a feminine one, particularly in the first part of the twentieth century (the view is based on Thibaudet terminology, adopted by N. Manolescu), in which one's external existence loses its importance and one's focus turns inward. The force and solid architecture of the Doric novel are replaced by vulnerability and the epic dilution specific to the Ionic novel. Here the inner conflict of/in the self no longer offers the same clarity as the character's opposition against with the world does. Therefore, meanings multiply, more often than not reaching contradictions and paradoxes, because one's inner, invisible world is far more difficult to explore than the outer one (Manolescu 1998). What is more, if up until the end of the nineteenth century biographical forms were deemed as *paraliterature*, from the interwar period onwards, the insertion of biographical and autobiographical elements consistently enriches epic matter, enhancing its veracity. The Doric or the transitive, on the one hand, and the Ionic or the reflexive, on the other, illustrate the growth forces of the novel, opposing forces respectively; according to the already classical terminology proposed by R.-M. Albérès, they also manifest as complementary forces, much like the two worlds, the visible and the invisible, that make up existence itself (Albérès 1968).

The third form, and perhaps the most complicated one, the Corinthian novel or the meta-novel summarizes previous formulas, while simultaneously distancing itself from them. It attests to a unique form of emancipation, which means that the novel is no longer tributary to mimesis, defying and surpassing the facts of reality and recognizable possibility. Furthermore, the meta-novel exhibits its functioning mechanisms, the writing itself becoming one of its fundamental themes, while the previous interrogation of the world and of the self is doubled or oftentimes supplanted by an interrogation of fiction itself. This is the final metamorphosis of the novel and,

naturally, the most spectacular one, as demonstrated by N. Manolescu in *Arca lui Noe* [Noah's Ark] (Manolescu 1998) or Radu G. Țeposu in *Viața și opiniile personajelor* [The Life and Opinions of Characters] (1983), as it represents the prerogative of great writers.

To what extent are these classifications valid for popular novels or for mass novels? Undoubtedly, the most frequent formula is the Doric one, even when dealing with romance novels, as the authors of mass literature do not seem to be prone to experiment with writing possibilities and are seldom concerned with the dynamic of narrative instances. Genre fiction thus includes novels of specific worlds and of certain social milieus; even when narration is done in the first person, they are not the novels of complicated individualities looking to understand the obscure mechanisms underlying existence. Knowledge in the abstract sense is of no concern to the protagonists of these fictional worlds and, we dare say, to the readers of these novels. The action in the popular novel is carried out *hic et nunc*, as it is a work of immanence, devoid of the feverish grand searches specific to the Ionic or Corinthian novels. Causality functions beyond reproach, the experienced reader suffers no bewilderment, while the evolution of the plot and the characters follows a predictable line. Of course, rudimentary psychological analysis and timid experimentation might appear, but technical awareness, the relationship between reality and fiction, the issue of veracity or authenticity do not bear the same weight as in the case of canonical works. The Ionic or the reflexive novel, and particularly the meta-novel require a specialized reader, able to conduct layered decoding, given their complicated referential function. The so-called triviality of events in the mass novel, the simplistic plot, and the sensationalism of such texts are incompatible with the narrative intricacies or with the character complexities. Their purpose is to cause enjoyment and to instruct, as their finality is an extraliterary one; accordingly, literariness is the result of plot and character construction, and less so of narrative or stylistic aspects.

The relevance and diversity of mass novels are self-evident, as we can still see in the bookstores' show windows, in supermarkets, airports, and train stations "detective novels, pop novels, Sci-Fi, 'pink' sentimental novels belonging to the wider category of the melodrama, erotica, sexy or pornographic novels, thrillers, and

horror novels from specialized collections such as *serie noire,* the lives of the famous, romanticized biographies of historical figures or, nowadays, of movie stars, pop stars or celebrities" (Drăgan 2006, 25).

Other classifications operate with thematic or main narrative technique antinomies: sentimental novels–cerebral novels (naturally, mass novels are sentimental, whereas modern, innovative novels are cerebral), family novels–single/ bachelor novels, monophonic/monologic novels–polyphonic/dialogic novels, subjective–objective, personal–impersonal, social–psychological, rural–urban, historical–mythical, realistic–sci-fi, novels–anti-novels, etc. If we consider the taxonomy proposed by novelist Anton Holban, mass novels are dynamic, constructed first and foremost on alert action, unlike the static novel, defined by analytical immovability (Holban 2005, 463–475). According to Jean Ricardou, on the one hand we have "the story of an adventure" (the definition of the classical novel as well as the mass novel), while on the other, "the adventure of the story" (the definition of the modern autotelic novel) (Ricardou 1988). The mass novel would thus correspond to the formula coined by René Girard: it is a romantic lie, unlike the romanesque truth of rare, canonical volumes, but, of course, genre purity is only possible on a theoretical level (Girard 1972). The novel of experience might seem as the modern descendant of the classical adventure novel; although the focus is mainly on inner adventures, much more dire than social ones, it otherwise contains the same ingredients: crises, the confrontation with death, betrayal, disappointment, etc. All these taxonomic doublets can be accompanied by more or less metaphorical definitions, particularly as we advance into the twentieth century: the black novel/ noir fiction, the "oceanographic" novel, Mircea Eliade's formula for the psychological novel (Eliade in Sasu and Vartic 1997, 175–176), "the novel of human condition" (Petrescu 1979), "the novel of the obsessive decade", "the nautical/ sea novel", etc. The novel of historical and/or personal trauma or the novel of memory (Holocaust novel, Gulag novel) is turning into one of the most widespread and validated current formulas, proving once again the strong connection between the genre and factuality. Obviously, the hosts of

possible classifications prove that we are indeed dealing with "a protean and genophage genre, meant to gloriously absorb other genres" (Drăgan 2006, 8).

Even if innovation is much more limited in the case of mass novels, here too an evolution that runs along with the evolution of human civilization is noticeable: from hajduks (outlaws) to superheroes, from adventure in the woods to outer space, the fictional universe of the common reader is molded alongside the lines of reality, and the varying combination of romanticism and realistic background remains the defining element for mass audience literature.

The Romance Novel of the nineteenth century, subsequently opposed by the cerebral novel of the Great Modernism, holds a predictable precedence: as the general public is predominantly made up of women, melodramatic plots, with their inter-relational thicket, are the most sought after. Don Juan figures, the *femme fatale(s)* or victims of villains populate this universe so prevalent nowadays in soap operas. The erotic novel is a variant of the romantic one, less common in the long nineteenth century (there are still ongoing lawsuits against writers on moral grounds, oftentimes caused by a confusion between reality and fiction), but with a posterity that confirms its impact in what social and individual emancipation is concerned.

The Hajduk/ Outlaw Novel is seemingly the most popular subgenre within the timeframe we are currently discussing about, which is to be expected as it heavily draws from folklore, on the one hand, but also from the Western model of the "Robin Hood" type of bandit, on the other. As a subgenre combining elements of the local imaginary with a foreign cultural influence, it demonstrates the way in which *paraliterature* is not entirely innocent in what text production mechanisms are concerned, and authors are not mere dilettantes, devoid of talent and culture. Like any strong representation of mentalities, the image of the hajduk was eventually confiscated by the interwar extreme right. However, after World War II, the prototype of the hajduk-hero will exert an influence on rock, folk, and pop music, countering nationalist suggestions and allowing for dissent. A variant of this type of novel is the bandit novel, in which the heroic features are belittled.

The Mystery Novel, which Marian Barbu defines as an *epiphenomenon* (Barbu 1981, 7), is in direct competition with the novel of everyday banality promoted by

High Realism. The imitation of Western models (e.g. *The Mysteries of Paris, The Mysteries of London*) is profoundly explicit even from the very titles, where only the toponymy is changed. Mysteries are, after all, incredible intrigues that draw, in turn, on the English Gothic novel of the eighteenth century, which, in order to contrast previous works from the Age of Enlightenment, had capitalized on the medieval imaginary. Another category partially overlapping *the mystery novel* is the *sensational novel* (this is, in fact, one of the most used paratextual labels), a formula that also emphasizes the importance of unrealistic, but extraordinary action, keeping readers on the edge of their seats. *The black novel* or *the criminal novel* (tags are based on their subtitles, meant to stir the buyer's curiosity) feeds the reader's imagination with spectacular intrigues, as epic augmentations of events found in newspaper publications, given the fact that some of the writers (such as Ilie Ighel) were first and foremost journalists.

Of all subgenres, the historical novel has the great advantage of not just a European tradition, but also of a local one because it builds upon historiographic writings, as well as upon folklore. The sword-and-sorcery novel, widespread in the West, has a direct correspondent in the Romanian context, chiefly when one has a look at the recipes of historical adventures. We can conclude, therefore, that "there are as many novel subgenres as there are media, crafts, techniques, and possible human situations" (Robert 1983, 55), and that the dominant Romanian social contexts of the nineteenth century have an impact on certain imaginary directions.

Researcher Ioana Drăgan proposes, in turn, a possible typology, adding to the subgenres mentioned above the novel of *the hero par excellence*, the novel of the *woman* or novels of adultery, revolt, individual heroism, judiciary or criminal novels (this is the nineteenth-century name for the detective novel), social-militant or apolitical novels (Drăgan 2006, 94). Beside this typological attempt, it is worth mentioning romanticized biographies, children's and young adult novels, the novel of the slums, travel novels. In certain cases, they are subdivisions (subgenres) of other formulas, but specific in their themes, chronotopes or intended audience.

An important erratum in the history of Romanian literature can be found in the case of the texts entitled *Catastihul amorului* [The Catastrophe of Love] and *La gura sobei* [By the Hearth], "the first writing which excellently parodies the two circuits, literary and paraliterary" (Drăgan 2006, 16) and "a literary anomaly in the context of the post-1848 Romanian novel, one of the literary oddities encountered in the beginnings of Romanian literature" (Drăgan 2006, 195). Published anonymously and later attributed to Radu Ionescu (Bălăieț in Ionescu 1974), the text proved to be only a localized translation (Teliban 2024). Even so, if we go back to the parodic exercise from *Țiganiada,* we might say that both the meta-discourse and the spirit of parody had been illustrated in the Romanian cultural context before the publication of the two much-debated pieces. However, given that Budai-Deleanu's work would only be published later, at the end of the nineteenth century, its echoes are feeble and not relevant for the evolution of the Romanian novel, which is the object of our chapter.

The Romanian novel in the digital age—archives, corpora, platforms

The digital market for literary texts, although initially timid (*Wikisource* and more or less clandestine PDF files), has started to progressively develop in recent years. Electronic books promoted by contemporary publishing houses are also a means to recover books of the past, difficult to render on paper. Digital libraries, particularly university ones, were first developed, followed by the successful attempt at creating corpora that would allow, apart from reading, other types of interventions on the texts. The first advancements were registered in the lexicology field, then the resources, methodologies, and tools gradually extended towards other philology areas.

A first example of an interoperable, open-access corpus, with novels encoded as TEI-xml, is the *European Literary Text Collection* [*ELTeC*]. This is a multilingual corpus, created with support from a COST action (*CA 16204: Distant Reading for European Literary History*) that started in 2017 and ended in 2022. Roxana Patraș was one of the MC*s* for the Romanian team and, as a part of WG1 and WG3, she coordinated the selection, digitalization, encoding, and analysis of the Romanian

collection (ELTeC–rom). Her role was also one of promoting Romanian literature as a valuable resource for cross-linguistic computational analyses. Each research group that participated in the project proposed an initial core selection of 100 novels for their respective national collections (to which a few more were subsequently added), based on the following criteria: genre diversity, varied text dimension (short, medium, long), reprint number (to balance out canonical and marginal texts). With a view to future stylometric experiments (which require both text variety and good sampling), some authors (between 9 and 11) were included with 3 books each, while others with only one item. The corpus entries were grouped in four periods (1840–1859, 1860–1879, 1880–1899, 1900–1920) in order to allow for a clearer observation of periodization issues, generational changes, and other serendipitous trends. The types of metadata included in the corpus refer to the author, the title, text composition characteristics, origin, first edition, language or script/ alphabet, aspects of orthography. In addition to being a fundamental resource, recovering many forgotten texts, the corpus has also generated relevant conclusions as a result of quantitative and comparative analyses. In its actual form (https://distantreading.github.io/ELTeC/rom/index.html), ELTeC–rom provides 100 Romanian novels published between 1840 and 1920 as well as a visualization on the relationship between texts written by men and those written by women and on the corpus' s time slots.

The Digital Museum of the Romanian Novel [*MDRR*] is the main result of a different project, coordinated by Ștefan Baghiu and Vlad Pojoga, with the support of the "Astra" Museum Complex in Sibiu, involving a large group of collaborators. At a first glance, perhaps *digital library* would have been a better-suited name; however, the reader-friendly interface and the visitors' access to covers of a bygone era offer the necessary justification for calling it a *museum* (https://revistatransilvania.ro/mdrr/, https://adt.arcanum.com/ro/collection/MuzeulDigitalAlRomanuluiRomanesc/). It is not just the contents, but also the very objects themselves that are brought back to the public's attention, who can also take note of the appearance of prints of old, rather than the digitized and processed texts. As it stands now, MDRR is an archive of the Romanian novel, encompassing three large eras: the nineteenth century

(with an index of series and first editions, including over 160 digitized novels, roughly 90% of the Romanian production of the age), the first part of the twentieth century, containing items with no copyright restrictions (1900–1932, over 360 digitized novels), and 1933–1947. The final period starts, therefore, with 1933, which Paul Cornea deemed "the golden year of the interwar Romanian novel", and ends with the first years of communism.

By using these resources and by corroborating qualitative and quantitative methods of analysis, researchers have conducted studies on the geography of the Romanian novel—either in the national context, or abroad—, on identity (social, ethnic, gender) or on the novel's internal temporality. Some conclusions may seem predictable, if we take into account the socio-historical context (for instance, the fact that in addition to Romanian, other frequent ethnic identities found in nineteenth-century novels are the Turks, Roma, and then the Jews). Nevertheless, the fact that the nineteenth-century Romanian novel is predominantly urban and not rural, as the famous theory by Lovinescu would have had us believe, was indeed a revelation (Baghiu *et al.* 2019).

Another accomplishment in terms of digitizing Romanian literature is a corpus of hajduk texts (Hai-Ro**)**, most of them novels and short prose texts presumably sharing the same topic and the same literary period (1850 and 1950). With a most fortunate acronym (referencing both Hajduk Romanian Novels, as well as the popular motto "Hai, România!" [Let's go, Romania!]), the corpus demonstrates that the hajduk subgenre was sufficiently widespread in mass culture, irrigated on the one hand by popular creations (ballads and hajduk suburban songs), and on the other hand, by the models of the Western heroic novel. Along authors specialized in this particular theme, such as N.D. Popescu or Panait Macri, many canonical authors, such as Mihail Sadoveanu or Panait Istrati, have capitalized on the hajduk prototype, and as such this strand has produced hypertexts and is worthy of in-depth research. Quantitative and qualitative methods of analysis have emphasized defining aspects for the Romanian hajduk context (the hajduk is a local picaro, crossing multiple suggestive spaces and media) or the specifics of the protagonists.

Coordinated by Roxana Patraș, the project *Romanul de consum și subgenurile sale în literatura română din secolul al XIX-lea: editare digitală și analiză de corpus literar asistate de calculator* [Romanian Popular Novels and Their Sub-genres during The Long Nineteenth Century: Digital Editing and Corpus Analysis]—shorthand Pop-Lite (*see* https://pop-lite.org)—is a continuation of the investigations carried out within *ELTeC* and *Hai-Ro* projects, which further recovers feuilleton novels, some even unfinished, suggestive of the way in which authors would quickly respond to public demands; the project, thus, indexes and samples not only edited volumes, but also some series and instalments. 123 samples are included in the corpus, novels or fragments of novels published between 1850 and 1920, belonging to best-selling authors of the long nineteenth century (N. D. Popescu, N. G. Rădulescu-Niger, Panait Macri, Ilie Ighel, George Baronzi, Theodor Stoenescu), as well as rarities signed by George Allan (Mite Kremnitz) or less-critically-acclaimed works belonging to canonical authors such as Vasile Alecsandri and B.P. Hasdeu. The interoperable format (TEI-xml), machine-readable and open-access, allows a diversified approach to texts, both with close-reading methods, as well as with complementary ones spawning from the distant reading approach, with diverse use cases in language history, lexicology, study of the imaginary or stylometry. Departing from the idea of operationalizing a "prototype" of the mass novel in the Romanian long nineteenth century, the selection criteria for the texts are: serialization, critical reception, republishing situation (some published in a single edition, others – best-sellers), text dimension, typeface, illustrations, type of paratext (whose scope was commercial, meant to attract readers), and, of course, the generic label (hajduk/ sensational/ romantic/ historical/ mystery/ criminal).

Another ongoing project is *MEMORO*: *Romanul memoriei în postcomunism: subgenuri, generații, rețele transnaționale* [The Novel of Memory in Post-Communism: Sub-genres, Generations, Transnational Networks], coordinated by Andreea Mironescu, whose chief importance lies in the fact that it indexes novels published after 1989, therefore it meritoriously touches on contemporary works (*see* https://www.novelofmemory.eu/acasă). As such, some of these projects have created

archives, corpora, and databases already available to the public and to researchers, while others are still ongoing and will offer an important amount of data and metadata.

The metaphor of the iceberg, suitable in genetic criticism, might, therefore, be extended at a macro-literary level. For the tip of the novel iceberg to exist, it must rely on an enormous, unseen mass, that would support it and offer it stability. Today's descendants of the mass novel developed in the nineteenth century—soap operas, internet mass literature, streaming series, comic books, graphic novels, etc.—demonstrate the vitality and relevance of this phenomenon, while complementary and interdisciplinary research methods ensure a wider coverage of the cultural milieu and the recovery of the digital cultural record.

References

Albérès, R. M. *Istoria romanului modern*. Traducere de Leonid Dimov. București: Editura pentru Literatură Universală, 1968.

Baghiu, Ștefan, *et al*. "Geografia internă a romanului românesc în secolul al XIX-lea." *Revista Transilvania*, nr. 10 (2019): 32. Accesat 12 mai 2020. https://revistatransilvania.ro/geografia-interna-a-romanului-romanesc-in-secolul-al-xix-lea/.

Barbu, Marian. *Romanul de mistere în literatura română*. Craiova: Scrisul Românesc, 1981.

Bodiu, Andrei. *Şapte teme ale romanului postpaşoptist*. Pitești: Paralela 45, 2002.

Burke, E. "Modernity's Histories: Rethinking the Long Nineteenth Century, 1750-1950." UC Berkeley: UC World History Workshop, 2000. Accesat la https://escholarship.org/uc/item/2k62f464.

Călinescu, G. *Istoria literaturii române de la origini până în prezent*. București: Minerva, 1986.

Cornea, Paul. *Regula jocului*. București: Eminescu, 1980.

Drăgan, Ioana. *Romanul popular în România*. Cluj-Napoca: Casa Cărții de Știință, 2006.

Eco, Umberto. *Opera deschisă*. Traducere de Cornel Mihai Ionescu. Pitești: Paralela 45, 2006.

Girard, René. *Minciună romantică şi adevăr romanesc*. Traducere de Alexandru Baciu. București: Univers, 1972.

Hobsbawm, Eric. *The Age of Extremes: The Short Twentieth Century, 1914–1991*. London: Abacusm (1995) [1994].

Holban, Anton. "Testament literar." În *Opere*, vol. I. București: Editura Fundației Naționale pentru Știință și Artă, 2005.

Ibrăileanu, G. "Creație și analiză." În *Scrieri alese*, ediție de A. Patraș și R. Patraș. Iași: Editura Universității „Alexandru Ioan Cuza", 2010.

Istrate, Ion (coord). *Dicționarul cronologic al romanului românesc de la origini până la 1989*. București: Editura Academiei Române, 2003.

Lovinescu, E. *Istoria literaturii române contemporane*, vol. 3. București: Minerva, 1981.

Manolescu, Nicolae. *Arca lui Noe. Eseu despre romanul românesc*. București: 100+1 GRAMAR, 1998.

Marino, Adrian. "Literatura de masă, Subliteratura, Paraliteratura." În *Biografia ideii de literatură*, vol. 5. Cluj-Napoca: Dacia, 1998.

Marino, Adrian. "Literatura populară, literatura de masă, paraliteratura, Heteronomia literaturii." În *Hermeneutica ideii de literatură*. Cluj-Napoca: Dacia, 1987.

Patraș, Roxana, Ioana Galleron, Lucreția Pascariu, Alexandra Olteanu, Ioana Lionte, and Camelia Grădinaru. *HAIRO: Corpus de Romane Haiduceşti (1850–1950)*, . 2020. https://www.nakala.fr/page/collection/11280/51a881d1.

Petrescu, Liviu. *Romanul condiției umane*. București: Minerva, 1979.
Petrescu, Liviu. *Vârstele romanului*. București: Eminescu, 1992.
Pillat, Dinu. "Cum se deschide la noi gustul pentru romanul senzațional." În *Mozaic istorico-literar. Secolul XX*. București: Eminescu, 1971.
Pillat, Dinu. "Romanul de senzație în literatura română din a doua jumătate a secolului al XIX-lea." În *Itinerarii istorico-literare*. București: Minerva, 1978.
Ricardou, Jean. *Noi probleme ale romanului*. Traducere de Liana și Valentin Atanasiu. București: Univers, 1988.
Robert, Marthe. *Romanul începuturilor și începuturile romanului*. Traducere de Paula Voicu-Dohotaru. București: Univers, 1983.
Sadoveanu, Mihail. *Amintiri literare*, antologie de Ion Bălu. București: Minerva, 1970.
Sasu, Aurel, și Mariana Vartic (eds). 1997. *Bătălia pentru roman*. București: ATOS.
Teliban, Andreea. 2024. "La curțile documentului. Două proze franțuzești în traducere românească: *Catastihul amorului* și *La gura sobei*." *Observator cultural*, nr. 1202, 4 aprilie 2024. https://www.observatorcultural.ro/articol/la-curtile-documentului-doua-proze-frantuzesti-in-traducere-romaneasca-catastihul-amorului-si-la-gura-sobei/.
Țeposu, Radu G. 1983. *Viața și opiniile personajelor*. București: Cartea Românească.

https://distantreading.github.io/ELTeC/rom/index.html
https://www.distant-reading.net/eltec/
https://poplite.org/
https://proiectulbrancusihairo.wordpress.com/
http://revistatransilvania.ro/mdrr
https://www.novelofmemory.eu/

Notes on the Contributors

Andrei-Marius AVRAM received a BS degree in computer engineering from Military Technical Academy of Bucharest "Ferdinand I" and a MS degree in Artificial Intelligence from University Politehnica of Bucharest. He is currently a machine learning engineer at Adobe and a PhD student at University Politehnica of Bucharest. In the past, he has been a research intern/data scientist at IBM, a research assistant at the Romanian Academy Research Institute for Artificial Intelligence (RACAI) and a senior machine learning engineer at Xperi. His research interests include machine learning, natural language processing, information retrieval and speech recognition.

Victoria BOBICEV is Associate Professor at the Technical University of Moldova, Faculty of Computers, Informatics and Microelectronics. Her main area of research is natural language processing and computational linguistics. The topic of her PhD thesis was statistical models of language and applied techniques of statistical language modeling with the aim of creating useful applications she investigated semantic features of text and the study of various phenomena in this field. She is committed to working with with students and entry level researchers with who she conducts various experiments on semantic features of languages, such as semantic relations between words and phonosemantics. She has also experimented with automatic sentiment analysis as an integral part of linguistic studies. In the last years she has been actively involved in the creation of the syntactically annotated corpus of the Romanian language within the Universal Dependecies project in which she is developingof the Romanian Non-Standard corpus. Currently she actively participates in the UniDive project dedicated to universal features and linguistic diversity from different parts of the world. Contact: victoria.bobicev@ia.utm.md

Tudor BUMBU holds a DSc since 2023 and is a junior researcher at the Institute of Mathematics and Computer Science of Moldova. His main areas of expertise are machine learning, NLP, data structures, artificial neural network, and natural language logic. He has coauthored several papers in impact journals and presented workshops and seminars in the digital preservation of the old Romanian heritage texts. Contact: tudor.bumbu@math.usm.md.

Liudmila BURTSEVA is ScD. in computer science, scientific researcher-coordinator at the Institute of Mathematics and Computer Science of Moldova State University. Her research interests include natural language processing, computational algebra, natural computing, and human-computer interaction. She published more than 80 scientific publications, including a monograph. Dr. Burtseva had participated in several national and international research projects. Her current research is performed in the field of digitization of Romanian historical printings and research software engineering. Contact: luburtseva@gmail.com

Maria CARP (MITROFAN) received her PhD degree in computer science from the School of Advanced Studies of the Romanian Academy, in 2020. She is currently a researcher at the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy. She was a

member of several NLP-related projects. Her research interests include natural language processing, deep learning, and biomedical NLP.

Mădălina CHITEZ is a senior researcher in applied corpus linguistics at the West University of Timișoara, Romania. She holds a doctoral degree from the Albert-Ludwig University of Freiburg, Germany, and worked or participated in research internships at several European universities (Oxford, Siena, Coventry, Zurich). Most of the research projects she has coordinated have resulted in free, user-friendly digital applications (e.g., LEMI). She is the founder and president of CODHUS (Centre for Corpus Related Digital Approaches to Humanities), which hosts multidisciplinary research with educational impact at the intersection of Corpus Linguistics, Digital Humanities, and Natural Language Processing. Email: madalina.chitez@e-uvt.ro

Georgeta CÎȘLARU holds a PhD in Linguistics and she is a Full Professor at University Sorbonne nouvelle. Her research areas are Text Linguistics, Writing, Semantics, Sociolinguistics, Discourse Analysis. Her recent research aims at elaborating a unified linguistic model to explain expectations, projections and predictions expressed through language. Its development involves integrating reception/understanding and production, articulating a transversal approach with the linearity of the linguistic flow, and observing expectations and interpretative failures in static, longitudinal and pathological contexts. She is the head of the Pro-TEXT ANR Project, co-authored *Le processus de textualisation* (2018, with T. Olive), edited Writing(s) at the Crossroads. The Process–product Interface (2015), Linguistic Approaches to Emotions in Context (2014) and contributed chapters on emotion in discourse. Contact: georgeta.cislaru@sorbonne-nouvelle.fr

Svetlana COJOCARU is a corresponding member of the Academy of Sciences of Moldova, professor, Dr. Hab. in computer science. She is vice-president of the Academy of Sciences of Moldova and principal scientific researcher at the Institute of Mathematics and Computer Science of Moldova State University. Her research interests include formal grammars and languages, natural language processing, computational algebra, and natural computing. She has about 300 scientific publications, including five monographs. Prof. Cojocaru has participated in several national and international research projects. The latest research is carried out in the field of digitization of Romanian historical printings. Contact: svetlana.cojocaru@math.md

Alexandru COLESNICOV is Dr. in computer science, scientific researcher-coordinator at the Institute of Mathematics and Computer Science of Moldova State University. His research interests include formal grammars and languages, natural language processing, computational algebra, natural computing, and computer typography. He published more than 150 scientific publications, including two monographs. Dr. Colesnicov had participated in several national and international research projects. His current research is performed in the field of digitization of Romanian historical printings. Contact: acolesnicov@gmx.com

Loredana CUZMICI holds a PhD in contemporary Romanian literature and she is a lecturer since 2015 at the Faculty of Letters, "Alexandru Ioan Cuza" University of Iași. She teaches lectures on Interwar Romanian prose, Literature, theater and film, and Research Methods in Humanities. She also coordinates the students' practice in teaching and applied philology. She authored a monograph *Generația Albatros - o nouă avangardă* (2015) and contributed several studies in collective volumes and journals. She coedited the digital literary corpus Pop-Lite (2024). Contact: loricuzmici@gmail.com

Ioana GALLERON is a Professor for French Literature and Digital Humanities at Université Sorbonne Nouvelle (Paris). Her main field of study is the French theater of the 17th and 18th century. She has published two monographs, nine collected studies, four critical editions and several digital editions, as well as numerous papers. She has been a chair of the COST action ENRESSH, and involved in several research projects financed by ANR, ANVUR, Campus France etc. She is currently a co-coordinator of the French consortium ARIANE (https://cst-ariane.huma-num.fr), part of HumaNum. Contact: ioana.galleron@sorbonne-nouvelle.fr

Radu ION joined the Romanian Academy Research Institute for Artificial Intelligence (RACAI) in 2001. He received his BSc in software engineering from Politehnica University of Bucharest in 2001, MSc in theoretical linguistics from University of Bucharest in 2003 and PhD in computer science from Romanian Academy in 2007. He is currently a senior research fellow at RACAI and his main research interests revolve around deep neural networks and deep learning for natural languages. Dr. Ion has published more than 100 papers in peer-reviewed journals and conferences and he is frequently invited to serve in program committees of many important conferences of the NLP field. Dr. Ion is currently a member of the editorial board of the Proceedings of the Romanian Academy Series A, the Information Science section.

Ludmila MALAHOV is a scientific researcher at the Institute of Mathematics and Computer Science of Moldova State University. Her research interests include formal grammars and languages, natural language processing, computational algebra, natural computing, and computer typography. She published more than 160 scientific publications, including three monographs. Dr. Malahov had participated in several national and international research projects. Her current research is performed in the field of digitization of Romanian historical printings. Contact: ludmila.malahov@math.md

Cătălina MĂRĂNDUC holds a PhD in philology and a PhD in computer science. Over the last years she has worked as a scientific researcher at the Institute of Linguistics "Iorgu Iordan - Al. Rosetti". Her fields of interest are general linguistics, computational linguistics, lexicography, pragmatics, pragmatics, stylistics, syntax, computational semantics and morphology, natural language processing tools, corpus linguistics. She actively participated in the creation and development of morphologically and syntactically annotated corpora of the Romanian language: the RoDia Romanian Diacronic and Romanian Non-Standard corpora within the Universal Dependencies project. Contact: catalina_maranduc@yahoo.com

Alexandra OLTEANU holds the position of Assistant Professor at *Alexandru Ioan Cuza* University of Iaşi. She is a member of the Digital Humanities Laboratory research team. She participated in the research teams of the ELTeC, Hai-Ro, and Pop-Lite projects. Her main scholarly interests are Cultural Studies, Digital Literary Studies, Genre Theory, Literary History and Theory. She has consistently researched the literature of the 19th century and the evolution of the Romanian novel. A selection of her articles comes as follows: *Modelul Gazetăresc al Banditului Social în Romanul Tâlhăresc al Secolului al XIX-lea* [The feuilleton model of the social bandit in the nineteenth-century outlaw novel], (2023); *The Ideological Instrumentalization of Literary Myths in 19th Century Romanian Literature* (2023); *Social Banditry and Feminity in the 19th Century Hajduk Novel,* (2023); *The Novel as the Foundation of Romanian Literary Terminology in the Age of Cultural Modernity* (2023). olteanualexandra99@yahoo.com.

Alexandru ORAVIȚAN is Junior Lecturer Ph.D. at the West University of Timișoara, Romania. He has conducted extensive research into post-9/11 literature as part of his doctoral work. Currently, his postdoctoral research endeavours revolve around contemporary American literature and culture, as part of his affiliation with the American Studies Center (CSAM), and applied linguistics, within the Centre for Corpus Related Digital Approaches to Humanities (CODHUS). He is a member of ESSE, EAAS, ELN, EUROCALL a.o. Email: alexandru.oravitan@e-uvt.ro

Vasile PĂIȘ received his BS from the Politehnica University of Bucharest, the Faculty of Automatic Control and Computers, then his PhD in computer science from the School of Advanced Studies of the Romanian Academy. He is currently a scientific researcher, in the field of artificial intelligence, working at the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy. He has been involved in many national and international research projects. His scientific papers were published in international journals, book chapters, and conference proceedings. His research interests include natural language understanding, language technology platforms and advanced neural network architectures.

Lucreția PASCARIU is a PhD student at "Alexandru Ioan Cuza" University of Iași with a thesis on the feuilleton-novel in nineteenth-century Romanian literature. She is a member of the Digital Humanities Laboratory (UAIC) and of the following project teams: Pop-Lite, Hai-Ro, ELTeC-rom, and Women Writers in History (CEEPUS). Her main areas of interest are cultural studies and digital literary studies. She has published articles about the structural function of the feuilleton-novel within the Romanian literary tradition and is one of the main contributors to the Romanian collection of ELTeC. Contact: lucretia.pascariu@gmail.com

Roxana PATRAS is Senior Researcher (I) at the Institute of Interdisciplinary Research, "Alexandru Ioan Cuza" University of Iași. She has been PI and member of several research projects funded by national and international grants. Her current research interests focus on modern Romanian and European literature, digital literary studies, literary theory and the development of open access instruments such as critical editions, dictionaries, anthologies, digital literary corpora, etc. She coordinated and (co)edited the first open access digital literary corpora sampling Romanian literature: ELTeC-rom (2022), Hai-Ro (2020), Pop-Lite (2024). Contact: roxana.patras@uaic.ro

Laura PRICOP holds a PhD in Philosophy since 2011 and she is scientific researcher III at the Institute of Interdisciplinary Research, "Alexandru Ioan Cuza" University of Iași. Her research is grounded in ethics, problematology and in philosophy of communication and analyses rationality structures in ethics and communication in the public sphere including doctor-patient communication. She has been science communication manager in the international research network CA16211: RECAST (2017-2021). Currently, she is an editor for *Open Philosophy* and deputy editor of the *Bulletin of Integrative Psychiatry*. Contact: laura.pricop@gmail.com

Voica PUȘCAȘIU is a lecturer on Modern and Contemporary Art History and Methodologies of Art History at Babeș-Bolyai University in Cluj-Napoca. Her research interests revolve around art in public spaces, both commissioned and unsanctioned for the analysis of which she employs a sociological perspective as she tries to establish how narratives are constructed and how viewers connect to artworks. More recently she's looking towards Memory Studies and issues of (mis)representation. As one of the founding members of DigiHUBB, Transylvania Digital Humanities Centre she is determined to introduce Digital Art History methodologies and tools in the

curriculum by focusing specifically on data visualisation and interactive cartography to open new paths of interpreting and teaching Art History. She is a co-trainer of the *Humanities Data and Mapping Environments* workshop at the European Summer University in Digital Humanities which she also helps organize alongside an interdisciplinary team at the Babeș-Bolyai University. Contact: voica.puscasiu@ubbcluj.ro

Matei-Alexandru STONESCU is an independent scholar with a PhD in Philology from CESI, University of Bucharest, specializing in Possible Worlds Theory. With a background in translation studies, suprasegmental phonetics, and spectrogram analysis, his research areas encompass modal logic, large language models and AI-driven NLP, network analysis, and data visualization. He has published several studies focusing on virtualization, modal truths, and Possible Worlds Theory. He served as the editor of the Romanian School of Rome journal "Ephemeris Dacoromana" from 2017 to 2022. Matei has published several studies focusing on virtualization, modal truths, and Possible Worlds Theory. Currently, he is using AI-driven analysis to trace consistency in public discourse and working on a project exploring the acoustic ecology of urban noise in Bucharest. Since 2019, he has been the Scientific Secretary of the Romanian Academy in Rome. Contact: matei.a.stoenescu@gmail.com

Elena Isabelle TAMBA is Senior Researcher I and coordinator of the Department of Lexicology and Lexicography, at the Institute of Romanian Philology "Alexandru Philippide", Romanian Academy-Iași Branch, where she is involved in the Lexical Thesaurus of the Romanian Language program, in two fundamental projects of the Romanian Academy (DLR. Romanian Language Dictionary. Litera C, and CLRE. Electronic Romanian Lexicographic Corpus) and in the eDTLR priority project. She has coordinated and participated in various research projects related to the digitization of lexicography or terminology. She is collaborating on the Dictionnaire étymologique des langues romanes (DÉRom). Her research areas are lexicography, e-lexicography, lexicology, lexicology, neology, terminology, Romanian grammar, Romanian history, Romanian philology and linguistics, computational linguistics, toponymy, and language policy. She has published books and articles in the country and abroad. Currently she is a delegate of the Romanian Academy at EFNIL https://efnil.org/. Contact: isabelle.tamba@gmail.com

Raluca TANASESCU is Researcher in Translation and Global Media in the School of Languages, Literatures, and Cultures at the University of Galway, Ireland. After obtaining a doctoral degree in Translation Studies from the University of Ottawa in Canada, she worked as a Postdoctoral Researcher in digital methods applied in early modern science history at the University of Groningen (the Netherlands). Her scholarly work is situated at the nexus of translation, complexity, and digital humanities. Raluca is mainly interested in the dynamics between translation and the digital as well as in the essential role played by the two in creating more sustainable and just communities. She is currently working on the European Research Council-funded project PIETRA: Translation and Communication and serves as the Chair of the Multilingualism and Multiculturalism Committee for the Alliance of Digital Humanities Organizations. Contact: raluca.tanasescu@universityofgalway.ie

Chris TANASESCU is a poet and academic with backgrounds in English and Computer Science and the author of *Literature and Computation: Platform Intermediality, Hermeneutic Modeling, and Analytical-Creative Approaches* (Routledge 2024). The Graph Poem project he started 15 years ago has outputted natural-language-processing and network-science-based poetry classifiers, intermedia performances, and computationally assembled poetry anthologies. His alias Margento refers to a

cyber cross-artform ensemble and international coalition of poets-translators, visual artists/ musicians, and coders throwing events and launching publications on and off-line in four continents. Chris is currently a Research Scientist on the PIETRA project at University of Galway. Previous or ongoing positions and affiliations include Coordinator of Digital Humanities at University of Ottawa, Altissia Chair in Digital Cultures and Ethics at Université Catholique de Louvain, Senior Researcher in Global Literary Studies and Complex Systems at Universitat Oberta de Catalunya, and Visiting Scholar at the Electronic Textual Cultures Lab, University of Victoria. He is an Asymptote Editor-at-Large. Contact: cristian.tanasescu@universityofgalway.ie

Dan TUFIȘ received his BS and PhD from the University Politehnica of Bucharest, the Faculty of Automatic Control and Computers. He is currently a full member of the Romanian Academy and was the Director of the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy, since 2002. Professor Dan Tufiș, published more than 300 scientific papers (books, book chapters, journal and conference papers) and delivered more than 90 invited talks all over the world. He coordinated or participated in more than 80 international and national research projects. He received various distinctions and prizes for his research throughout his career and is currently Doctor Honoris Causa of the University Agora of Oradea and Honorary professor of the University "A.I. Cuza" of Iași. He received the highest presidential order "Steaua României" and the "Nicolae Milescu Spataru Medal, awarded by the Academy of Republic of Moldova. He is a member of the Editorial Boards of several international journals. His research interests include natural language processing, machine translation, knowledge representation, semantic web.