

**Mihai-Bogdan Atanasiu • Anca-Diana Bibiri •  
Emanuel Grosu • Alina Moroşanu •  
Constantin Răchită**  
(Editors)

## **CULTURAL DYNAMICS OF VALUES**



EDITURA UNIVERSITĂȚII „ALEXANDRU IOAN CUZA” DIN IAȘI

---

2024

The book of proceedings of the conference  
*Interdisciplinary Perspectives in Humanities and Social Sciences*  
9<sup>th</sup> Edition: Rethinking Values in Interdisciplinary Research (27-28 October 2023)  
Institute of Interdisciplinary Research  
Department of Social Sciences and Humanities  
"Alexandru Ioan Cuza" University of Iasi

**Scientific Reviewers:**

Prof. Adina Dornean, PhD habil.

Senior Researcher Marius Chelcu, PhD habil.

**Language revision of texts in English:**

Senior Researcher Anca-Diana Bibiri, PhD

**Book editor:**

Emanuel Grosu

Marius-Nicușor Grigore

**Coverage:** Manuela Oboroceanu

ISBN: 978-606-714-906-7

DOI: 10.47743/phss-2024

The authors are entirely responsible for the scientific contents of the texts herein published as well as for the fair use of the copyrighted works.

© Editura Universității „Alexandru Ioan Cuza” din Iași

700539 – Iași, Str. Munteni nr. 34, tel. (0232) 314947; editura@uaic.ro

www.editura.uaic.ro

**Mihai-Bogdan Atanasiu** is a senior researcher, director of the Department of Social Sciences and Humanities, Institute of Interdisciplinary Research, “Alexandru Ioan Cuza” University of Iași. He has a PhD in History, awarded in 2012 at the same university. His research activity focuses on the political, social and cultural history of Moldavia in the seventeenth and eighteenth centuries. Most of his scholarly contributions have focused on social history, genealogy, prosopography, history of the Church, history of mentalities, as well as on editing documentary sources.

**Anca-Diana Bibiri** is a senior researcher at the Department of Social Sciences and Humanities, Institute of Interdisciplinary Research, “Alexandru Ioan Cuza” University of Iași. She has a PhD in Philology, and a postdoctoral fellowship in linguistics from the same University, and her main areas of research are: prosody, phonetics and dialectology, computational linguistics, natural language processing, lexicography, and sociolinguistics. Co-editor of the PHSS Proceedings (2014-2019).

**Emanuel Grosu** is a senior researcher at the Department of Social Sciences and Humanities, Institute of Interdisciplinary Research, “Alexandru Ioan Cuza” University of Iași. He has a PhD in Philology. He published studies, exegesis and translations from medieval Latin authors (Paulus Diaconus, Dungalus Reclusus, Anselm of Canterbury, Marcus of Regensburg, Marco Polo), the diachronic evolution of central literary themes and motifs of medieval Latin culture constituting the main research direction. Co-editor of the PHSS Proceedings (2014-2019).

**Alina Moroșanu** is a senior researcher at the Department of Social Sciences and Humanities, Institute of Interdisciplinary Research, “Alexandru Ioan Cuza” University of Iași. She has a PhD in in Cybernetics and Economic Statistics, awarded in 2011 at the same university. Her research interest include: questionnaires development, healthcare management analytics, project management, statistical analysis, statistical software (R, SPSS), surveys, human resources analytics.

**Constantin Răchită** is a research assistant in the Department of Social Sciences and Humanities, Institute of Interdisciplinary Research, “Alexandru Ioan Cuza” University of Iași. He has a PhD in Philology. His primary research focuses on the translation and interpretation of ancient texts in Old Greek and Latin. Throughout his career, he has participated in various translation and editing projects. Currently, his research interests encompass interdisciplinary approaches to biblical and patristic texts, exploring issues related to translation, transmission, and their influence on contemporary society.

**Descrierea CIP a Bibliotecii Naționale a României**

**Cultural Dynamics of Values** / Mihai-Bogdan Atanasiu, Anca-Diana

Bibiri, Emanuel Grosu, .... - Iași : Editura Universității “Al. I. Cuza”, 2024

Conține bibliografie

ISBN 978-606-714-906-7

I. Atanasiu, Mihai-Bogdan

II. Bibiri, Anca-Diana

III. Grosu, Emanuel

## Contents

<b>Foreword</b> ( <i>Anca-Diana Bibiri</i> ) .....	9
--	---

### Plenary Conference

Human Enhancement: tehnologie versus teologie. Repere pentru o evaluare interdisciplinară a valorilor și posibilităților de devenire a umanului prin cunoaștere [Human Enhancement: Technology versus Theology. Landmarks for the Interdisciplinarity Evaluation of Human Values a Potential of Becoming Through Knowledge]

<b>Pr. Andrei-Răzvan Ionescu</b> .....	21
--	----

### Philology

The Use of Artificial Intelligence (AI) in Linguistics. Case Study: Analysis of Linguistic Phenomena in the Novel *Ion* by Liviu Rebreanu

<b>Cristina Bleorțu</b> .....	35
-------------------------------	----

Traducerea automată a literaturii. O himeră încă vie? [Automatic Translation of Literature: A Still Living Chimera?]

<b>Alexandra Ilie</b> .....	51
-----------------------------	----

Kitsch. The Control and Faking of Aesthetic Value

<b>Daniela Petroșel</b> .....	77
-------------------------------	----

Valori perene în predarea romanisticii în spațiul universitar românesc [Perennial Values in Teaching Romance Studies in Romanian Universities]

<b>Mihaela Secieru</b> .....	93
------------------------------	----

Authentic vs. Pseudo Values

<b>Paula-Andreea Onofrei</b> .....	111
------------------------------------	-----

Medical Humanities Approached Through a Feminist Lens

<b>Laura Ioana Leon</b> .....	127
-------------------------------	-----

Explorări teoretice și suprapuneri terminologice. Romanul, obiect de reflecție și prim suport al teoriei genurilor [Theoretical Explorations and Terminological Overlaps. The Novel, as Object of Reflection and the First Support of the Genre Theory]

<b>Alexandra Olteanu</b> .....	141
--------------------------------	-----

Spectrele filiațiilor literare. Portrete ale generațiilor – Mircea Ivănescu și Radu Vancu [The Specters of Literary Filiations. Portraits of Generations – Mircea Ivănescu and Radu Vancu]	
<b>Teodora Iurusiuc</b> .....	165
Memoria comunismului în <i>Jurnalul unui jurnalist fără jurnal de Ion D. Sîrbu</i> [The Memory of Communism in Ion D. Sîrbu's <i>Journal of a Journalist without a Journal</i> ]	
<b>Oana-Elena Nechita</b> .....	181
Language in the Church: Orthodox Religious Terminology in Polish and the Role of Translations in Establishing Lexical Norms	
<b>Irina-Marinela Deftu</b> .....	201
<b>History &amp; Theology</b>	
<i>Non naturalibus desideriiis, sed censibus aestimentur.</i> Piața romană de legume și fructe [ <i>Non naturalibus desideriiis, sed censibus aestimentur.</i> The Roman Vegetable and Fruit Market]	
<b>Iulia Dumitrache</b> .....	219
Un posibil clivaj între teoria asupra stilului și aplicarea acesteia în cazul Fericitului Augustin? Pluralitatea de stiluri în operele acestuia [A Possible Split Between the Theory of Style and Its Application in the Case of Augustine? The Plurality of Styles in Augustine's Works]	
<b>Pr. Liviu Petcu</b> .....	249
Conflicting Values during the French Wars of Religion (1562-1598): Loyalty to the King and Loyalty to God	
<b>Andrei Constantin Sălăvăstru</b> .....	261
Polemici teologice în <i>Praefatio paraenetica</i> a lui John Pearson (1613-1686) [Theological Polemics in <i>Praefatio paraenetica</i> by John Pearson (1613-1686)]	
<b>Constantin Răchită</b> .....	281
Moartea – o preocupare a vieții cotidiene în Iași veacului al XVIII-lea [Death – a Preoccupation of Everyday Life in the 18 <sup>th</sup> Century Iași]	
<b>Mihai-Bogdan Atanasiu</b> .....	301

## Economics

The Use of Information and Communication Technologies in Business as a Value-Creating Tool: Analysis on the Case of SMEs in Romania <b>Valentina Diana Rusu &amp; Angela Roman</b> .....	317
Evaluation of Hospital Financing in Romania: A Comparative Analysis pre- and post-Pandemic COVID-19 <b>Mihai-Vasile Pruteanu &amp; Alina Moroşanu</b> .....	337
Green Jobs, Green Skills and Green Human Resource Management. An Analysis of Current Trends <b>Silvia-Maria Carp &amp; Ana-Maria Bercu</b> .....	367
Is Security a Timeless Value? An Insight from International Relations <b>Andreea-Cosmina Foca &amp; Oana-Maria Cozma</b> .....	381

# The Use of Artificial Intelligence (AI) in Linguistics.

## Case Study: Analysis of Linguistic Phenomena in the Novel *Ion* by Liviu Rebreanu

CRISTINA BLEORȚU\*

**Abstract:** *In recent years, artificial intelligence (AI) has made significant progress, particularly in the field of natural language processing (NLP). This growth has only accelerated since the COVID-19 pandemic, which prompted a surge in the use of digital tools and remote technologies. The rapid development of AI has greatly enhanced NLP capabilities, allowing for the swift creation of concordances, identification of collocates, and generation of frequency counts of linguistic features. Tasks that once took hours or even days can now be completed in just seconds with a few lines of code. These advances not only improve efficiency but also reduce the likelihood of human error, eliminating the need for labour-intensive processes such as manually sifting through large Excel documents. Furthermore, AI-driven tools have become essential in linguistic research, facilitating more complex analyses of texts, including syntactic parsing, sentiment analysis, and semantic interpretation. This paper aims to demonstrate the practical application of AI tools, specifically those developed for the Romanian language by Professor Dan Tufiş's team (Institute of Artificial Intelligence, Romanian Academy), for the analysis of the novel *Ion* by Liviu Rebreanu. By leveraging the power of AI, we seek to provide deeper insights into the linguistic features of the text. To conduct this analysis, we will use the R programming language and tools such as `udpipe`, which offer robust methods for text processing and linguistic annotation. This study not only showcases the potential of AI in humanities research but also highlights the growing importance of digital tools in advancing the study of the Romanian language.*

**Keywords:** *Artificial Intelligence; Linguistics; Natural Language Processing; Romanian language; Programming language.*

---

\* Scientific Researcher, PhD, Faculty of Letters and Communication Sciences, "Ștefan cel Mare" University of Suceava, Romania; [cbleortu@hotmail.com](mailto:cbleortu@hotmail.com).

## Introduction

In the past, linguistic research relied heavily on traditional methods, often involving painstaking manual work with pen and paper. Analyzing linguistic features such as syntax, morphology, and semantics was a time-consuming process, limited by the available resources and the sheer effort required to manually sift through large volumes of text. However, with the advent of the digital age, artificial intelligence (AI) and big data have revolutionized the field, enabling the rapid and automated analysis of vast amounts of linguistic data. These technologies have become essential for identifying new patterns and trends in language, transforming the massive influx of data generated daily into a rich linguistic resource for the 21<sup>st</sup> century. Tasks that once demanded thousands of hours of manual effort can now be executed with remarkable efficiency and precision, often at the click of a button.

Previously, researchers were often constrained by small-scale surveys or limited experiments to explore linguistic phenomena, as the tools and data necessary for more extensive research were not readily available. Today, technological advances have dramatically expanded these possibilities. Large corpora containing millions of words, such as those accessible through platforms like Sketch Engine<sup>1</sup>, are now routinely employed for comprehensive linguistic studies. These corpora enable a wide range of analyses, from basic frequency counts to more complex investigations into collocations, discourse patterns, and sociolinguistic variations.

Moreover, sophisticated linguistic algorithms now facilitate the analysis of millions of data points. Automated transcription tools like Whisper<sup>2</sup> and Microsoft Word<sup>3</sup> allow researchers to quickly transcribe interviews and spoken language data, while annotation tools provide detailed morphosyntactic and lexical information. These capabilities, which were previously out of reach, have opened up new avenues for linguistic research, allowing for more in-depth and nuanced studies of language use

---

<sup>1</sup> <https://www.sketchengine.eu/> [16.09.2024].

<sup>2</sup> <https://github.com/openai/whisper> [16.09.2024].

<sup>3</sup> Only the web version.



across different contexts and genres. Additionally, researchers can now create huge corpora by extracting data from the internet through web-scraping techniques. This has greatly expanded the scope and depth of linguistic analysis, providing a wealth of new data for exploring linguistic phenomena in diverse languages and contexts:

*But it is common now for corpora to range from 1 billion words, like the GeoWAC family of corpora (Dunn & Adams, 2020), up to 400 billion words, like the Corpus of Global Language Use (Dunn, 2020). These very large corpora are often drawn from digital sources like the web, social media, Wikipedia, and news articles. While these sources of language data have tremendous potential for testing linguistic hypotheses on a large scale, working with them requires computational methods to scale up the analysis. [...] Computational models allow us to scale up our analysis to very large corpora. This allows linguists to analyze an amount of data that, as individuals, we could never hope to analyze<sup>4</sup>.*

Modern technologies in speech recognition, synthesis, and translation are transforming the field of linguistics, enabling more detailed and complex analysis of languages. Advanced artificial intelligence (AI) systems now play a crucial role in this transformation by learning and predicting user preferences through sophisticated algorithms, which tailor and personalize content on various platforms. The integration of AI in online interactions not only enhances user experiences but also contributes to linguistic research by offering tools that can process and analyze large datasets of spoken and written language with high accuracy<sup>5</sup>.

In addition to AI, the use of cloud computing has further accelerated these advances. Cloud-based systems provide the necessary computational power and storage capacity to handle vast amounts of linguistic data, facilitating real-time processing and analysis. This has opened up new possibilities for linguistic research, allowing for the exploration of language patterns, structures, and usage on an unprecedented scale.

---

<sup>4</sup> Jonathan Dunn, *Natural Language Processing for Corpus Linguistics*, Cambridge University Press, Cambridge, p. 3 and p. 12.

<sup>5</sup> Cristina Bleorțu, Lavinia Seiciuc, *Big data și inteligența artificială în cercetarea lingvistică*, Casa Cărții de Știință, Cluj-Napoca, 2024, p. 10.

These technological developments have also significantly impacted the humanities, especially in the realm of corpus linguistics. The adoption of standard markup languages such as HTML (HyperText Markup Language), XML (Extensible Markup Language), and TEI (Text Encoding Initiative<sup>6</sup>) has enabled researchers to annotate and encode linguistic data systematically. Furthermore, standard platforms like Open Journal Systems<sup>7</sup> have streamlined the creation and distribution of academic journals, enhancing the accessibility and dissemination of scholarly work.

The shift towards a more technologically driven approach in linguistics began to gain momentum in the mid-20<sup>th</sup> century. It marked the beginning of a more structured and systematic method for linguistic analysis and the digital humanities. This era saw the convergence of traditional linguistic study with computational methods, leading to the development of tools and frameworks that support extensive linguistic research. Today, these advances continue to shape the future of linguistics, enabling researchers to explore and understand languages in ways that were once thought impossible:

„Utilizarea<sup>8</sup> calculatoarelor în științele umaniste a început în anii 1950, odată cu dezvoltarea traducerii automate, marcând un moment crucial în integrarea tehnologiilor informaționale în studiile umaniste<sup>9</sup>. De la jumătatea secolului al XX-lea, științele umaniste digitale au cunoscut o expansiune semnificativă, depășind granițele Europei și Americii de Nord, unde au fost inițial concepute, și extinzându-se la nivel mondial datorită procesului de globalizare, care a facilitat conectarea între oameni, culturi și idei diverse (Fiormonte 2023: 19). Această evoluție a fost marcată de o interacțiune productivă între specialiștii în informatică și cercetătorii în domeniul științelor umaniste, colaborare care reprezintă atât o oportunitate

---

<sup>6</sup> <https://tei-c.org/> [16.09.2024].

<sup>7</sup> <https://openjournalsystems.com/> [16.09.2024].

<sup>8</sup> Cristina Bleorțu, Lavinia Seiciuc, *Noi metode de cercetare în studiile umaniste: edițiile digitale*, Casa Cărții de Știință, Cluj-Napoca, 2024, p. 27.

<sup>9</sup> S. Schwandt (ed), *Digital Methods in the Humanities: Challenges, Ideas, Perspectives* (Digital Humanities Research, 1), Bielefeld University Press, Bielefeld, 2021, p. 7. <https://doi.org/10.14361/9783839454190>

majoră, cât și o provocare semnificativă pentru domeniul emergent al umanioarelor digitale<sup>10</sup>”.

However, despite the significant global growth of digital humanities, the field is still in its infancy in Romania<sup>11</sup>. Researchers identify the official emergence of digital humanities in the country having originated circa 2014. Since then, most studies in Romania have centred on showcasing digital tools or reviewing foreign publications rather than actively applying these tools to Romanian-language texts.

The situation is similar within corpus linguistics, a key subfield of digital humanities. Compared to other major Romance languages and languages such as English and German, Romanian has a noticeably smaller amount of resources and research devoted to building comprehensive linguistic corpora. This lack of resources has limited the ability to analyze Romanian language patterns and structures on a large scale, underscoring the need for more focused research efforts in this area.

„În România<sup>12</sup>, însă, lingvistica de corpus a cunoscut o dezvoltare și mai întârziată. Proiecte majore precum *Corpusul Lexicografic Românesc*

---

<sup>10</sup> [‘The use of computers in the humanities began in the 1950s with the development of machine translation, marking a crucial moment in the integration of information technologies into humanistic studies (*ibidem*). Since the mid-20th century, digital humanities have experienced significant expansion, surpassing the boundaries of Europe and North America—where they were initially conceived—and extending globally due to the process of globalization, which has facilitated connections between people, cultures, and diverse ideas (Domenico Fiormonte). This evolution has been marked by a productive interaction between computer science specialists and researchers in the humanities, a collaboration that represents both a major opportunity and a significant challenge for the emerging field of digital humanities.’, own translation].

<sup>11</sup> We must also acknowledge the contributions of the two centers in Iași and Suceava – namely, the Institute of Interdisciplinary Research (ICI) and the Interdisciplinary Center for Humanities and Arts. These institutions are poised to play a crucial role in advancing this field, contributing significantly to the development of digital humanities and corpus linguistics in Romania. Their interdisciplinary approach and commitment to fostering research in the humanities are essential for expanding the resources and knowledge base related to the Romanian language.

<sup>12</sup> Cristina Bleorțu, Lavinia Seiciuc, *Noi metode de cercetare...*, pp. 36-37.

*Electronic (CLRE)* și *CoRoLa* au fost demarate abia în 2014, marcând începutul unei noi ere în cercetarea lingvistică românească. De asemenea, au fost obținute corpusuri importante prin tehnici moderne, cum ar fi *RoTenTen*, care au contribuit la extinderea resurselor lingvistice pentru studiul limbii române. Cu toate acestea, în contrast cu alte limbi romanice, România nu dispune încă de un corpus istoric complet finalizat. Lipsa unui astfel de corpus reprezintă o limitare pentru cercetătorii interesați de studierea evoluției istorice a limbii române, împiedicând o analiză diacronică detaliată și sistematică. În prezent, eforturile sunt orientate spre dezvoltarea și completarea unui astfel de corpus, care ar constitui un instrument esențial pentru avansarea cunoașterii în domeniul lingvisticii istorice românești<sup>13</sup>.

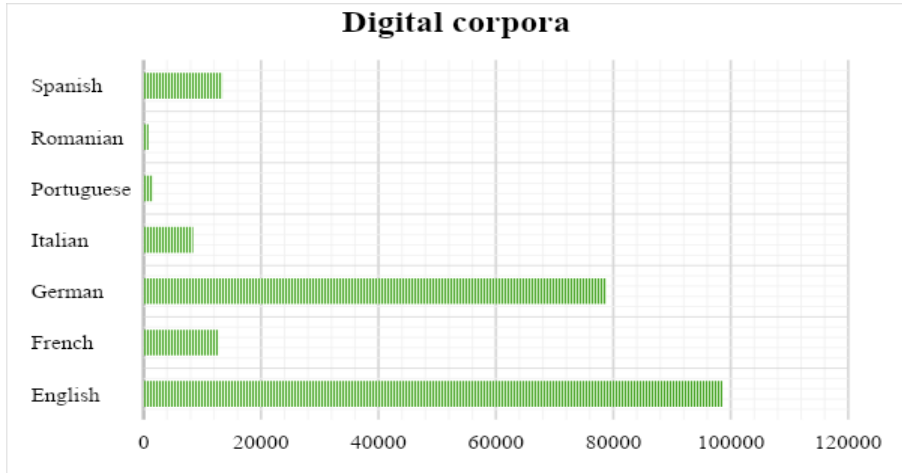
A glance at the *Virtual Language Observatory*<sup>14</sup> confirms this disparity. Romanian has access to significantly fewer digital corpora compared to other languages. While major languages like English, German, and the dominant/the most widely used Romance languages offer a vast array of comprehensive corpora, Romanian lags behind with a limited selection of resources for linguistic research. This shortage underscores the urgent need for more targeted efforts to develop digital linguistic resources for the Romanian language, to enable more thorough and diverse linguistic analysis<sup>15</sup>:

---

<sup>13</sup> [‘In Romania, however, corpus linguistics has developed at an even slower pace. Major projects like the Romanian Electronic Lexicographic Corpus (CLRE) and CoRoLa were initiated only in 2014, marking the beginning of a new era in Romanian linguistic research. Additionally, significant corpora have been obtained through modern techniques, such as RoTenTen, which have contributed to expanding linguistic resources for the study of the Romanian language. Despite these advances, unlike other Romance languages, Romania still lacks a fully completed historical corpus. The absence of such a corpus poses a limitation for researchers interested in studying the historical evolution of the Romanian language, preventing detailed and systematic diachronic analysis. Currently, efforts are focused on developing and completing such a corpus, which would serve as an essential tool for advancing knowledge in the field of Romanian historical linguistics.’, own translation].

<sup>14</sup> <https://www.clarin.eu/content/virtual-language-observatory-vlo> [16.09.2024].

<sup>15</sup> This does not mean that the existing corpora, such as CoRoLa and RoTenTen, are not important or that the efforts invested in their creation are not valuable. On the

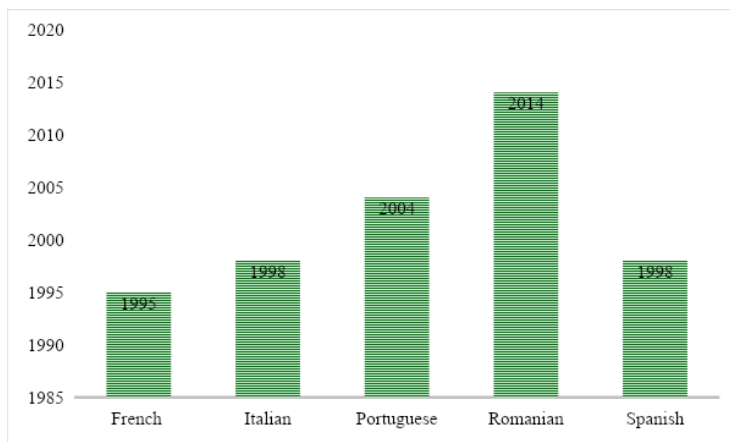


**Figure 1.** Digital Corpora in Major Romance Languages, German, and English at the Virtual Language Observatory

If we examine the timeline of when the first digital corpus was published for the major Romance languages, a similar pattern emerges: for Romanian, this milestone occurred somewhat later than for the other primary Romance languages. While languages such as Spanish, French, and Italian established their first digital corpora relatively early, for Romanian, this development came later, further contributing to the current gap in linguistic resources and research efforts between Romanian and other more prominent languages. This delay has had a lasting impact on the availability and development of comprehensive linguistic tools and databases for Romanian:

---

contrary, these corpora represent significant strides in advancing linguistic research for the Romanian language. They provide crucial foundations upon which further studies can be built, and the work done so far is commendable for paving the way toward a more comprehensive understanding of Romanian linguistics.



**Figure 2.** The Evolution of Corpus Linguistics in the Romance Languages

As is evident, there remains much to be accomplished in the field of digital humanities, particularly regarding the Romanian language. To address this, we will demonstrate how one can contribute to this domain by utilizing tools developed by Professor Tufiş's group, along with a program like R, and applying them to a well-known Romanian novel, *Ion* by Liviu Rebreanu. This approach will showcase how these resources can be used to enrich linguistic analysis and corpus development, helping to bridge the current gap in digital humanities for the Romanian language.

### **Analysis of the Novel *Ion* by Liviu Rebreanu**

To analyze the aforementioned novel using the R programming language, several steps need to be undertaken: (1) preprocessing the text, (2) standardizing the corpus, and (3) extracting linguistic phenomena of interest from the novel *Ion*.

The first step is to download the R programming language<sup>16</sup> and the RStudio integrated development environment<sup>17</sup>; R is essential for performing statistical analysis and data processing, while RStudio provides a user-friendly interface for writing and executing R code.

<sup>16</sup> <https://www.r-project.org/> [16.09.2024].

<sup>17</sup> <https://posit.co/download/rstudio-desktop/> [16.09.2024].

## Preprocessing the text

To preprocess the text, we first need a web page where the novel is published. Then, we must install the 'pdftools' package and library in R, which allows us to extract text from PDF files for further analysis.

```
install.packages("pdftools")  
library(pdftools)
```

To use a corpus in PDF format, the first step is to input the URL where the novel is located. This allows the text of the novel to be downloaded and converted into a format suitable for analysis. Once the URL is provided, the novel can be extracted, saved as a PDF file, and then processed for further analysis using the steps outlined previously. This method is particularly useful for integrating digital texts into a corpus not only for linguistic studies but also for literary studies:

```
url <- "https://bgrmihailsturdza.wordpress.com/wp-  
content/uploads/2014/02/rebreanu-liviu-ion-i-carte.pdf"
```

To download and use the text from the PDF file, we need to use another function from the same pdf tools package in R. Specifically, we can use the pdf\_text() function, which allows us to extract the text content from the PDF file for further analysis:

```
pdf_text <-  
pdf_text(URL)
```

To simplify the processing and analysis of the text, you can convert it into a single block of text using the paste() function in R. This function concatenates the text from all pages into one continuous string, making it easier to perform further text analysis:

```
full_text <- paste(pdf_text, collapse  
= " ")
```

## Standardization of the corpus

Before analyzing a text, it needs to be properly prepared to ensure accurate and meaningful results. This preparation involves standardizing the text to eliminate variations that could affect the analysis. A key part of this process is converting all characters to lowercase, which prevents issues arising from differences in capitalization. Additionally, punctuation marks should be removed, as they are typically not relevant for most forms of text analysis and can introduce noise into the data.

In R, these tasks can be accomplished with the following functions:

```
full_text <- to lower(full_text)
full_text <- gsub("[[:punct:]]", "", full_text)
```

## Splitting the Text into Words

Another crucial step to facilitate text analysis is splitting the corpus into individual words. This process, known as tokenization, breaks down the continuous text into separate tokens, typically words. By doing this, we can analyze the frequency, structure, and usage of words within the text more effectively.

In R, this can be achieved using functions such as `strsplit()`:

```
words <- unlist(strsplit(full_text,
"\s+"))
```

**Strsplit()** splits the text into pieces based on a specified delimiter. In this case, the delimiter is `\s+`, which represents one or more whitespace characters. This ensures that the text is split wherever there is a space, creating a list of individual words.

**Unlist()** is used to convert this list into a simple vector of words for easier analysis.

By splitting the text into words, we prepare the data for more detailed analyses, such as word frequency counts, keyword extraction, and other lexical analyses that can provide deeper insights into the content and style of the text.



In the case of our novel, this process results in a total of 82,735 words:

Values	
full_text	" liviu\nrebreanu\n ion\n \n colecie iniati i coor...
pdf_text	Large character (251 elements, 524.7 kB)
url	"https://bormihailsturdza.wordpress.com/wp-content/upload
words	Large character (82735 elements, 1.4 MB)

**Figure 3.** The environment of R and the number of words from the novel *Ion*

This tokenization provides a foundational dataset for further linguistic analysis, allowing us to explore various aspects of the novel and thematic patterns. By working with this large collection of individual words, we can uncover deeper insights into the author’s language use and stylistic choices.

### Extracting Verbs from the Corpus

If, for example, we want to automatically extract verbs from the text, we need to go through a few additional steps. First, we need to install a package that allows R to identify the parts of speech within the text. For this purpose, we will use the tool specifically created for the Romanian language by Professor Tufiş’s group: `udpipe`, which is a powerful AI tool for linguistic processing that provides part-of-speech tagging, lemmatization, and dependency parsing for various languages, including Romanian. By using this tool, we can identify and extract verbs and other grammatical categories directly from the corpus.

```
install.packages("udpipe")
library(udpipe)
```

After loading the package and the library, the next step is to load the model for the Romanian language:

```
model <- udpipe_download_model(language =
"Romanian")
ud_model <- udpipe_load_model(model$file_model)
```

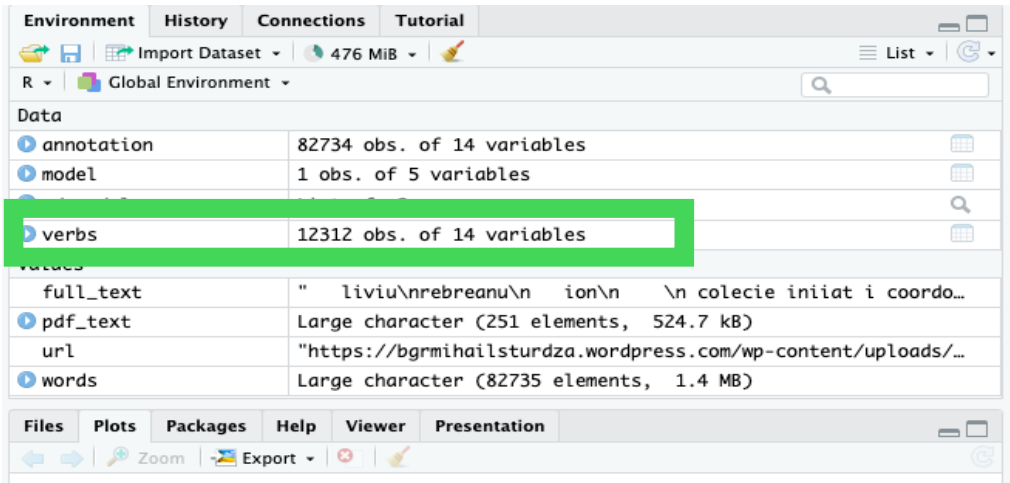
Finally, using the same `udpipe` package, we can annotate the text and extract the verbs. This process involves applying the loaded Romanian language model to the text, tagging each word with its part of speech, and then filtering out the verbs for further analysis:

```

annotation <- udpipe_annotate(ud_model, x =
full_text)
annotation <- as.data.frame(annotation)
verbs <- subset(annotation, upos == "VERB")

```

After executing these functions, more than 12,000 verbs are obtained, which can be analyzed:



Object	Details
annotation	82734 obs. of 14 variables
model	1 obs. of 5 variables
verbs	12312 obs. of 14 variables
full_text	" liviu\nrebreanu\n ion\n \n colecie iniat i coordo...
pdf_text	Large character (251 elements, 524.7 kB)
url	"https://bgrmihailsturdza.wordpress.com/wp-content/uploads/...
words	Large character (82735 elements, 1.4 MB)

**Figure 4.** The environment of R and the number of verbs from the novel *Ion*

These are just a few examples of the possibilities that R and AI tools like `udpipe` offer for linguistic analysis. By leveraging these tools, researchers can perform complex tasks such as part-of-speech tagging, lemmatization, and syntactic analysis with ease. This opens up a wide range of opportunities for exploring language patterns, understanding text structures, and gaining deeper insights into both literary and non-literary texts. With the growing integration of AI in linguistic research, tools like these are becoming increasingly vital for advancing the field of digital humanities and corpus linguistics.

## Conclusions

Given that there is still significant room for growth in the field of digital humanities in Romania, especially in corpus linguistics, this paper seeks to demonstrate how we can contribute to its development using AI tools. In particular, we utilized resources developed by the group led by Dan Tufiş at the Mihai Drăgănescu Institute of Artificial Intelligence in Bucharest, where Professor Tufiş serves as director. His team's work has been instrumental in advancing computational linguistics for the Romanian language, offering tools that facilitate a range of linguistic analyses.

We began by providing an overview of digital humanities, the importance of big data, and the current state of Romanian corpus linguistics. This overview highlighted the gaps and opportunities for research in this area, emphasizing the need for more digital corpora and analytical tools tailored to the Romanian language. To address these gaps, we then moved to a practical example, using the novel *Ion* by Liviu Rebreanu. This classic work of Romanian literature served as a case study to show how linguistic analysis can be performed using the R programming language and tools such as `udpipe`.

Through R, we demonstrated various stages of text analysis, including text preprocessing, annotation, and extraction of linguistic features such as verbs. The R programming environment, enhanced by packages like `udpipe`, allows researchers to process large texts efficiently – transforming them into structured data for analysis. While we only scratched the surface of what is possible, these examples underscore the versatility and power of R in linguistic research. From basic text cleaning and tokenization to more advanced tasks like part-of-speech tagging, R provides a robust toolkit for researchers.

Our intention was not just to provide a technical demonstration but to inspire further exploration and encourage more extensive use of these tools in Romanian linguistic research. By utilizing these AI-powered tools, researchers can conduct more profound and nuanced analyses, enhancing our understanding of the Romanian language and literature. We hope this discussion will pique interest in the potential of digital humanities and corpus linguistics, showing that even complex analyses can be carried out with relative ease and minimal technical expertise.

Moreover, the integration of AI in digital humanities represents a transformative shift in how we approach linguistic studies. The use of computational tools not only accelerates the analysis process but also opens up new avenues for research that were previously inaccessible. For instance, the ability to analyze large volumes of text quickly contributes to a deeper understanding of language evolution, stylistic variations, etc. By continuing to leverage these technologies, we can advance the field of Romanian corpus linguistics, ensuring that it remains at the forefront of digital humanities research. Ultimately, our work aims to encourage scholars, educators, and students to embrace these tools, supporting a more dynamic and inclusive research environment.

## Bibliography

- Bleorțu, C., Cuevas-Alonso, M., "Inteligencia artificial y análisis de rasgos lingüísticos en corpus de textos híbridos. El caso del castellano y el asturiano", *Dialectologia*, 2024 (forthcoming).
- Bleorțu, C., Cuevas-Alonso, M., Villazón Valbuena, M., Lamar Prieto, C., "Lingüística e inteligencia artificial. El corpus de La Pola Siero, Star Scholars Press", Star Scholar, 2024 (forthcoming).
- Bleorțu, C., Seiciuc, L., *Big data și inteligența artificială în cercetarea lingvistică*, Casa Cărții de Știință, Cluj-Napoca, 2024.
- Bleorțu, C., Seiciuc, L., *Noi metode de cercetare în studiile umaniste: edițiile digitale*, Casa Cărții de Știință, Cluj-Napoca, 2024.
- Dumitrescu, Ș. D., Boroș, T. și Tufiș, D., "RA-CAI's Natural Language Processing pipeline for Universal Dependencies", *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 174-181, Vancouver, 3-4.08.2017:  
<http://universaldependencies.org/conll17/proceedings/> [16.09.2024], 2017.
- Dunn, J., *Natural Language Processing for Corpus Linguistic*, Cambridge University Press, Cambridge, 2024.
- Eshkol-Taravella, I., Lefevre-Halftermeyer, A., "Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage", *Corela*. <http://corela.re-vues.org/4800> [archive]; 10.4000/corela.4800 [06. 08.2024], 2017.
- Patraș, R., Galleron, I., Grădinaru, C., Lionte, I., Pascaru, L., „The Splendors and Mist(eries) of Romanian Digital Literary Studies: a Stat-of-Art just before Horizons 2020 closes off”, *Hermeneia. Journal of Hermeneutics, Art Theory and Criticism*, 23, 2019, pp. 209-222.

- Păiș, V., Radu, I. Avram, A.-M., Mitrofan, M. și Tufiș, D., "In-depth evaluation of Romanian natural language pipelines", *Romanian Journal of Information Science and Technology*, vol. 24, 4, 2021, pp. 384-401.
- Rojo, G., *Análisis informatizado de textos*, Universidad Santiago de Compostela, Santiago de Compostela, 2023.
- Rojo, G., *Introducción a la lingüística de corpus en español*, Routledge, New York, 2021.
- Rumsey, D. J., *Statistics for Dummies*, 2nd edition, John Wiley & Sons Inc, New York, 2016.
- Sadin, É., *La humanidad aumentada. La administración digital del mundo*, Caja Negra, Buenos Aires, 2017.
- Schwandt, S. (Ed.). *Digital Methods in the Humanities: Challenges, Ideas, Perspectives* (Digital Humanities Research, 1), Bielefeld University Press, Bielefeld, 2021.  
<https://doi.org/10.14361/9783839454190>
- Winter, B., *Statistics for Linguists. An Introduction Using R*, Routledge, New York, 2020.

### **Corpus**

Liviu Rebreanu, *Ion*, Litera Internațional, București, Chișinău, 2020.