

Editors:

Adriana Baci (Birluțiu)

Daniela Nagy-Onița

Dan Tufiș

Dan Cristea

LINGUISTIC RESOURCES AND TOOLS FOR NATURAL LANGUAGE PROCESSING

Proceedings of the 19th International Conference,
Alba Iulia, 14-16 November 2024



Editura Universității “Alexandru Ioan Cuza” din Iași
2025

Organisers:

“1 Decembrie 1918” University of Alba Iulia

Institute for Computer Science, Romanian Academy, Iași branch

Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

“Mihai Drăgănescu” Research Institute for Artificial Intelligence,
Bucharest

“Alexandru Philippide” Institute of Romanian Philology, Romanian
Academy, Iași branch

Romanian Association of Computational Linguistics

Academy of Technical Sciences

Under the auspices of the Romanian Academy

This volume was published with the support of
the Faculty of Computer Science,
“Alexandru Ioan Cuza” University of Iași

ISSN: 3045-2074

ISSN-L: 3045-2074

ISSN (online): 3045-2538

PROGRAM COMMITTEE

Vasile Apopei, Romanian Academy, Iași branch – Institute of Computer Science

Adriana Baci (**Birluțiu**), “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Verginica Barbu Mititelu, Research Institute for Artificial Intelligence, Romanian Academy

Victoria Bobicev, Technical University of Moldova

Mihaela Colhon, Faculty of Mathematics and Natural Sciences, University of Craiova

Svetlana Cojocaru, Academy of Sciences of Moldova, Institute of Mathematics and Computer Sciences

Dan Cristea, Romanian Academy, Iași branch – Institute of Computer Science

Corina Forăscu, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Daniela Gifu, Romanian Academy, Iași branch – Institute of Computer Science and “Alexandru Ioan Cuza,” University of Iași - Faculty of Computer Science

Gabriela Haja, Romanian Academy, Iași branch – “Alexandru Philippide” Institute of Romanian Philology

Adrian Iftene, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Diana Zaiu Inkpen, School of Electrical Engineering and Computer Science, University of Ottawa, Canada

Radu Ion, Research Institute for Artificial Intelligence, Romanian Academy

Elena Irimia, Research Institute for Artificial Intelligence, Romanian Academy

Mihai Alex Moruz, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Daniela Marcela Nagy-Onița, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Constantin Orasan, University of Surrey

Vasile Florian Păiș, Research Institute for Artificial Intelligence, Romanian Academy

Mircea Petic, Vladimir Andrunachievici Institute of Mathematics and Computer Science Moldova State University and ACETI

Andrei Scutelnicu, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science and Romanian Academy, Iași branch - Institute of Computer Science

Elena Isabelle Tamba, Romanian Academy, Iași branch – “Alexandru Philippide” Institute of Romanian Philology

Horia-Nicolai Teodorescu, Romanian Academy, Iași branch – Institute of Computer Science and “Gheorghe Asachi” Technical University of Iași – Faculty of Electronics, Telecommunications and Information Technology

Diana Trandabăț, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Ștefan Trăușan-Matu, National University of Science and Technology Politehnica Bucharest – Faculty of Automation, Control and Computer Engineering and Romanian Academy - Research Institute for Artificial Intelligence

Dan Tufiș, Romanian Academy – Research Institute for Artificial Intelligence

Anca Vasilescu, Transylvania University of Brașov – Department of Mathematics and Computer Science

ORGANISING COMMITTEE

Mihaela Aldea, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Vasile Apopei, Romanian Academy, Iași branch – Institute of Computer Science

Adriana Baci (**Birluțiu**), “1 Decembrie 1918” University of Alba Iulia - Faculty of Informatics and Engineering

Cecilia Bolea, Romanian Academy, Iași branch – Institute of Computer Science

Nicoleta Breaz, “1 Decembrie 1918” University of Alba Iulia - Faculty of Informatics and Engineering

Matei-Vasile Căpîlnaș, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Dan Cristea, Romanian Academy, Iași branch - Institute of Computer Science

Daniela Cristea, “1 Decembrie 1918” University of Alba Iulia - Faculty of Informatics and Engineering

Corina Forăscu, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Lucian Gădioi, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Daniela Gîfu, Romanian Academy, Iași branch - Institute of Computer Science and “Alexandru Ioan Cuza” University of Iași - Faculty of Computer Science

Adrian Iftene, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Manuela Kadar, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Maria Muntean, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Daniela Marcela Nagy-Onița, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Lucian Popa, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Corina Rotar, “1 Decembrie 1918” University of Alba Iulia – Faculty of Informatics and Engineering

Andrei Scutelnicu, Romanian Academy, Iași branch – Institute of Computer Science and “Alexandru Ioan Cuza” University of Iași - Faculty of Computer Science

Diana Trandabăț, “Alexandru Ioan Cuza” University of Iași – Faculty of Computer Science

Dan Tufiș, Romanian Academy – Research Institute for Artificial Intelligence

TABLE OF CONTENTS

FOREWORD.....	VII
KEYNOTE SPEECHES.....	1

CHAPTER 1. DIGITAL HERITAGE

DIGITALISATION OF MEDIEVAL HANDWRITTEN TEXTS	7
TUDOR BUMBU, LIUDMILA BURTSEVA, MARIUS CERESCU, SVETLANA COJOCARU, ALEXANDRU COLESNICOV, LUDMILA MALAHOV	
NEW ACHIEVEMENTS IN THE DIGITISATION OF LINGUISTIC ATLASES OF THE ROMANIAN LANGUAGE	23
SILVIU-IOAN BEJINARIU, LUDMILA MALAHOV, TUDOR BUMBU, FLORIN-TEODOR OLARIU, RAMONA LUCA	

CHAPTER 2. CORPORA

ROMEMES: A MULTIMODAL MEME CORPUS FOR THE ROMANIAN LANGUAGE	45
VASILE PĂIȘ, SARA NIȚĂ, ALEXANDRU-IULIUS JERPELEA, LUCA PANĂ, ERIC CUREA	
DO FAIRY TALES INFLUENCE THE ACQUISITION OF LANGUAGE IN PRIMARY SCHOOL? STEPS TOWARDS AN AUTOMATIC EVALUATION.....	64
MARIA UNGUREANU, ALEX MORUZ, LAURENȚIU ȘOITU, DAN CRISTEA	
EXPLORING LEGAL LINGUISTIC PATTERNS: A CORPUS-BASED ANALYSIS OF ROMANIAN LAW TEXTS.....	79
ALEXANDRU DINU	
ROMANIA IN A MULTILINGUAL CORPUS OF EUROPEAN POETRY.....	95
SILVIE CINKOVÁ, PETR PLECHÁČ	
HUMAN VERSUS AI-REWRITTEN ACADEMIC WRITING: A CORPUS-BASED COMPARATIVE ANALYSIS OF PHRASEOLOGICAL COMPLEXITY IN ESSAYS AND SCIENTIFIC ARTICLES	107
MĂDĂLINA CHITEZ, ROXANA ROGOBETE, ANA-MARIA BUCUR, ANDREEA DINCA, IOANA-SIMINA FRÎNCU	

CHAPTER 3. APPLICATIONS

REDDIT IS ALL YOU NEED: AUTHORSHIP PROFILING FOR ROMANIAN	125
ECATERINA ȘTEFĂNESCU, ALEXANDRU-IULIUS JERPELEA	
A MULTIDIMENSIONAL APPROACH FOR TEXT ANALYSIS	138
VLAD BELCIUG, ELENA PELICAN	

ENHANCING PATIENT COMPREHENSION OF MEDICAL REPORTS THROUGH LLM-BASED LANGUAGE SIMPLIFICATION: THE CASE OF MEDX.....	154
VALENTIN NAGACEVSCHI, TUDOR P. TOMA	

CHAPTER 4. LARGE LANGUAGE MODELS

EXPLORING LARGE LANGUAGE MODELS FOR TRANSLATING ROMANIAN COMPUTATIONAL PROBLEMS INTO ENGLISH	171
ADRIAN MARIUS DUMITRAN, ADRIAN-CĂTĂLIN BADEA, ȘTEFAN-GABRIEL MUSCALU, ANGELA-LILIANA DUMITRAN, ȘTEFAN-COSMIN DĂSCĂLESCU, RADU-SEBASTIAN AMARIE	
DETECTING FAKE NEWS USING BERT AND ROBERTA MODELS: A RESEARCH STUDY	188
ASMII SHRIPAD	
ARE GENERATIVE PRE-TRAINED TRANSFORMERS ALL WE CAN DO IF WE WANT TO HAVE A SMART MACHINE?	203
DAN CRISTEA, DANIELA GÎFU, ANDREI SCUTELNICU, CECILIA BOLEA	
THE DISCOURSE OF PROFESSIONAL COMMUNICATION WITH YOUNG LEARNERS: CAREER ORIENTATION PRACTICES IN THE FIELD OF COMPUTATIONAL LINGUISTICS	221
ANDREEA NECHIFOR, MARIA UNGUREANU	
INDEX OF AUTHORS.....	239

FOREWORD

The volume includes the papers of the 19th edition of the international Conference “Linguistic Resources and Tools for Natural Language Processing”, which was hosted, this time, by Alba Iulia, a wonderful city located in the center of Romania. From its inception, in 2001, the ConsILR Conference (the acronym comes from *Consortiul de Informatizare pentru Limba Română – Consortium for Informatisation of the Romanian Language*), an initiative born in the Section of Information Science and Technology of the Romanian Academy, was meant as a fix point for linguists and computational linguists, but also for PhD students and master students in Humanities and Computational Linguistics. The initial focus of the ConsILR series of conferences has been the study of the Romanian language from a computational perspective. In time, the view of the organisers largely broadened, making ConsILR not being strictly addressed to researchers working on Romanian language but also to other scientists, who could thus share their experiences in modelling and implementing techniques for various other languages. Today, many modern language technologies have a universalist perspective, with generic solutions applicable to more than one language (see, for instance, the past Universal Dependencies project, or the recent COST action UniDive, accounting for universality and diversity-driven development of language resources and tools).

The host organiser of this edition was University “1 Decembrie 1918” of Alba Iulia, with its Department of Mathematics and Computer Science, which worked in close connection with colleagues from the Faculty of Computer Science of the “Alexandru Ioan Cuza” University in Iași, the Romanian Academy, with three institutes: the “Mihai Drăgănescu” Research Institute for Artificial Intelligence in Bucharest, the Institute of Computer Science and the “Alexandru Philippide” Institute of Romanian Philology, the last two located in Iași, as well as with the Romanian Association of Computational Linguistics and the Academy of Technical Sciences of Romania. An important contribution was made by the members of the Scientific Committee, thanks to whose reviewing efforts the quality of the papers was substantially improved, and we take this opportunity to thank them for their assistance.

As always along its rather long history, this edition of the Conference also tried to keep up with the incredible fast developments in the domain. The recent advances of language technology based on large language models (LLMs) could not be overlooked, making the keynote speeches and more regular papers report relevant aspects of this mainstream research.

Four keynote speeches were presented during the Conference, all having been very well appreciated by the audience, for bringing well documented field news, mainly fed from inside the groups of researchers led by the speakers

themselves. These prominent researchers and the titles of their speeches (with biographical short notes and abstracts included in this volume) are:

- Marius Leordeanu – **Into the Reality of Mind: from Imagination to Artificial Intelligence**,
- Andrei Păun – **Computing with Artificial Neural Networks: SNP systems**,
- Diana Inkpen – **Explainable Emotion and Sentiment Classification**,
- Verginica Barbu Mititelu – **Making Hay While the Sun Shines**.

The blind reviews selected 13 papers (from the 20 submitted) for being presented in the Conference and included in this volume. They are grouped in four sections:

- **Digital Heritage**,
- **Corpora**,
- **Applications**, and
- **Large Language Models**.

The two papers in the **Digital Heritage** section, deal with issues of extremely high importance for the preservation of Romanian cultural heritage: the digitization process of medieval handwritten texts and of linguistic atlases.

Five papers were grouped in the **Corpora** section. An interesting paper is dedicated to a curated dataset of real memes in the Romanian language, with multiple annotation levels. By exploiting a digitised corpus of fairy tales and a collection of compositions commissioned to pupils, the second study proves the positive impact in the development of the linguistic performance of children enrolled in the first-cycle school after being exposed to fairy tales. The following paper deals with a corpus of legal texts, investigating the construction and analysis of a representative legal corpus for the Romanian language and comparing it with a general literary corpus. A very interesting study reveals how Romanian geo-names are reflected in a multilingual corpus of European poetry. Finally, phraseological concepts are used as investigation tools to detect the threat of generative AI technologies in the elaboration of scientific texts by humans.

Three papers are grouped in the **Applications** section. It is no surprise that all of them are based on LLM approaches: an original view of an authorship profiling approach based on the Reddit social media platform, a multidimensional design of a text analysis system that provides a comprehensive and interpretable understanding of text characteristics, and an interesting presentation of the design and evaluation of an interface that, using the AI generative technology, bridges the communication gap between medical professionals and patients, by converting complex medical terminology into simple language, comprehensible to individuals without medical training.

FOREWORD

The most technical section is specifically dedicated to **Large Language Models** and also includes three papers. One study shows that robust LLMs, when given well-structured prompts, can reliably translate from Romanian into English less common texts, as in the technical jargon of the International Olympiad in Informatics. Then, the issue of fake news and their accurate detection is the topic of the next paper. Finally, the last paper in the section analyses and discusses the strengths and weaknesses of generative pre-trained transformer models with respect to tasks that have been until recently considered problematic for these systems.

In this edition of ConsILR, the organisers had the initiative to dedicate a special session to high school young students, in order to raise their interest for this fascinating field of science, so much linked to both sides of school didactic activities, humanities and exact sciences. As such, the last paper in the volume resumes the interventions in this special session.

To conclude, this volume offers new and valuable contributions to the development of the field of natural language processing, not few of them being focused on applications for the Romanian language. Several issues are present-day concerns of our scientific community, irrespective of the investigated language. We believe that this volume will be a very useful instrument for NLP researchers, as well as for students preparing their MSc dissertations or PhD theses in the area of computational studies of natural languages.

March 2025
The Editors

KEYNOTE SPEECHES

MARIUS LEORDEANU

Short bio: Marius Leordeanu is Professor of Computer Science at the National University of Science and Technology Politehnica Bucharest and Senior Researcher at the Institute of Mathematics of the Romanian Academy (IMAR). His research aims to push further our understanding of intelligence, study its ways of connecting to the world and construct our reality of mind. He received a bachelor degree in Computer Science and Mathematics from the City University of New York (2003) and a PhD from the Robotics Institute of Carnegie Mellon University (2009). At Politehnica he introduced in 2014, for the first time in the Computer Science Department, courses on Computer Vision and Robotics. At IMAR he started in 2016 the Advanced Computer Vision Seminar, with weekly meetings. For his work on unsupervised learning for graph matching, he received in 2014, as a single recipient, the prestigious award “Grigore Moisil” offered by the Romanian Academy. In 2020, his book *Unsupervised Learning in Space and Time* appeared in Springer, quickly becoming one of the best-selling books on the topic of unsupervised learning. At the *2024 Romanian Research Gala*, he received the Grand Prize in Computer Science and Mathematics for his scientific results obtained throughout his career together with his group of skillful young researchers and doctoral students.

Besides science, he is interested in music and literature. He composed and co-produced a professional album and a single (*Supersonic* – 2022, *Cosmic Attraction* – 2024). He writes poetry (with three volumes already published: *Povestea unui Cuvânt* – 2013, *Odiseu și Destinul* – 2024, and *Ființele soarelui*, at *Curtea Veche Publishing House* – 2024), same as prose (with one book published, *Mă numesc albastru* – 2016). In his activities he always aims to connect, in a natural and coherent way, both from human and computational perspectives, different domains of human experience, in order to better understand the real world, at interplay between mind and matter.

Contact: <https://sites.google.com/site/mariusleordeanu/>,
marius.leordeanu@cs.pub.ro

Title: Into the Reality of Mind – from Imagination to Artificial Intelligence

Abstract: What does it mean “to see”? Vision takes a lot of our cognitive power, relating all levels of our understanding of the world, from seeing objects and their actions and interactions, to understanding higher-level and more abstract events, to visualizing full stories in space and time. How can vision represent such stories and what is the connection between vision and language? Is there a common and explicit representation between the two? In his talk, Marius presented his recent

work on trying to answer this essential question, into the more general realm of Human-AI interaction – to better understand ourselves and how we can best use AI to extend our capabilities and to transform our imagination into reality.

ANDREI PĂUN

Short bio: Andrei received his BSc in Mathematics with a major in Computer Science from University of Bucharest. He then received a Master's followed by a PhD degree in Computer Science from the University of Western Ontario (Canada), in 2003. His PhD supervisor was Dr. Sheng Yu, from whom he “inherited” the interest on Finite Automata and OOP modelling as well as the Biocomputing research field, which was the main area of his PhD dissertation. Following the PhD, Andrei was a postdoc at the Technical University of Wien (Austria) and later at University of Rovira i Virgili (Spain). He then moved to the US as an assistant and, later, associate professor of Computer Science (tenured) at Louisiana Tech University. Later, he returned to Europe having positions at University of Bucharest and the National Institute of RandD for Biological Sciences. Recently he accepted a position of director of the “Mihai Drăganescu” Research Institute for Artificial Intelligence of the Romanian Academy. His main interests are Biocomputing, Bioinformatics, AI as well as finite automata and other related areas.

Contact: <https://www.racai.ro/en/about-us/racai-staff/apaun>; apaun@racai.ro

Title: Computing with Artificial Neural Networks – SNP systems

Abstract: Following the recent Nobel prize awards in Physics and Chemistry, Andrei provided a novel view of the AI field, as well as other models of computing inspired from the working of neurons: Spiking Neural Systems (SNP). The SNP systems are a type of distributed, parallel, neural-like computation model inspired by how biological neurons process information and communicate through spikes. In the past decade, our understanding of neurons and the brain has advanced significantly, largely due to the Human Brain Project in Europe and the BRAIN Initiative in the US. This discussion will highlight findings related to SNP systems and their variants as computational models, many demonstrating Turing Universality under certain constraints. He also examined recent research and ongoing questions, particularly concerning Spiking Neural P systems with communication on request (SNQ P systems). Notably, he has designed compact, Turing universal SNQ P systems with a limited number of neurons, including a generative SNQ P system that performs acceptance and function computation using just 12 neurons (11 for acceptance and 21 for function computing).

DIANA INKPEN

Short bio: Diana Inkpen is a Professor at the University of Ottawa, in the School of Electrical Engineering and Computer Science. She received her Ph.D. in Computer Science from the University of Toronto, Canada, and her M.Sc. and B.Eng. in Computer Science and Engineering from the Technical University of Cluj-Napoca, Romania. Her research is in applications of Natural Language Processing and Deep Learning. She is the editor-in-chief of the *Computational Intelligence* journal and the associate editor for the *Natural Language Engineering* journal. She published a book on Natural Language Processing for social media (Morgan and Claypool Publishers, *Synthesis Lectures on Human Language Technologies*, the third edition appeared in 2020), 11 book chapters, more than 40 journal articles, and more than 140 conference papers. She received many research grants, from which the majority include intensive industrial collaborations.

Contact: <https://www.site.uottawa.ca/~diana/>; diana.inkpen@uottawa.ca

Title: Explainable Emotion and Sentiment Classification

Abstract: Deep learning classifiers based on Large Language Models (LLMs) can achieve high performance, but they are black boxes: their decisions are not explainable. Users cannot fully trust these decisions. We present experiments in generating explanations with the LLMs, and we compare them with baseline explanations produced by older methods such as LIME, SHAP, and Integrated Gradients. We build several state-of-the-art classification models on datasets such as the IMDB movie review database (2 classes) and the SemEval 2018 Emotion dataset (11 emotions classes, where more than one label is allowed for a text). Then we generated human-understandable explanations for model predictions based on prompt engineering, and evaluating them with automatic metrics such as sufficiency, and with human judges (in terms of relevance, understandability, and trustworthiness). We show that for both evaluations, the generative method produces better explanations than LIME, SHAP, and the Integrated Gradients method.

VERGINICA BARBU MITITELU

Short bio: Verginica Barbu Mititelu is a senior researcher in the Natural Language Technology group of the Romanian Academy Research Institute for Artificial Intelligence and representative of the *Multiword Expression* section in the SIGLEX board. She performed her Master studies and received her PhD in Philology in 2010 from the University of Bucharest. She has constantly been preoccupied with and involved in the development of language resources, especially for Romanian, applying up-to-date annotation schemas and adjusting

them to the characteristics of the language under study. She has also been concerned with standardising the resources developed, especially using Linked Data principles of representation, and with the registration of their metadata in international data repositories.

Contact: <https://www.racai.ro/about-us/racai-staff/verginica-barbu-mititelu/>;
vergi@racai.ro

Title: Making hay while the sun shines

Abstract: Call it an expression, a locution, an idiom, a construction, etc., word combinations that show idiosyncrasy at any linguistic level are a pervasive language phenomenon, which has long been coined as a “bottleneck” for Natural Language Processing. The proliferation of large language models in today’s language technology is salutary, though the processing of the above-mentioned word combination is not a solved problem yet and tasks such as machine translation are still hindered by their occurrence. In her talk Verginica presented the challenges raised by this phenomenon, some language resources that have been developed for such Romanian word combinations, all against the background of the effervescent international community preoccupied with its description and treatment in language processing.

CHAPTER 1

DIGITAL HERITAGE

DIGITALISATION OF MEDIEVAL HANDWRITTEN TEXTS

TUDOR BUMBU, LIUDMILA BURTSEVA, MARIUS CERESCU, SVETLANA
COJOCARU, ALEXANDRU COLESNICOV, LUDMILA MALAHOV

*Moldova State University, “V. Andrunachievici” Institute of Mathematics and
Computer Science, Chisinau, Republic of Moldova*

bumbutudor10@gmail.com

luburtseva@gmail.com

marius.cerescu@gmail.com

svetlana.cojocaru@math.md

acolesnicov@gmx.com

ludmila.malahov@math.md

Abstract

Compared to recognizing printed texts, handwritten ones present a higher degree of difficulty. The paper examines approaches for the recognition of uncials writing, the dataset being based on three Romanian manuscripts written in uncial Cyrillic script from the 16th century: *Floarea darurilor*, *Codex Sturdzanus*, and *Codicele Bratul*. The problems in processing these historical documents lie in the variety of handwriting styles, the lack of dictionaries and well-defined grammatical rules appropriate to the period in question, the continuous writing of the text (without separating words and sentences with spaces or other punctuation marks), etc. ABBYY FineReader, versions 15 and 16, was used as a tool for OCR. The paper describes the development of specialized models and presents their accuracy concerning the volume of the dataset, which, in the best cases, reaches 90% even with relatively small training datasets. The possibilities of using AI applications for OCR and for split-ting text into words and sentences are studied.

Keywords: medieval manuscript digitalisation, handwritten text recognition, Romanian uncial script, OCR, *scripta continua* (continuous writing).

1. Introduction

The digitalisation of cultural heritage has become increasingly important with the advancement of in-formation technology. We have been working on digitising and transliterating Romanian texts printed in Cyrillic characters from the 17th to the 20th centuries, leading to the creation of the HeDy platform for processing historical texts (Bumbu et al., 2023).

We aim to broaden our study by focusing on medieval manuscripts written in uncial Cyrillic script. These manuscripts, dating from the 13th to the mid-19th

century, are available in significant numbers. However, studying them is challenging because there are few digital copies, and they need to be trans-literated into modern Latin script to make them accessible for us to work with.

Recognizing handwritten texts is a much more difficult problem compared to printed texts processing, as handwriting is an individual characteristic of the person. Furthermore, in the case of old handwriting, differences can also be observed in texts written by the same person, even within one and the same document.

One of the major challenges for early medieval texts is the continuous script (*scripta continua*), which lacks spaces or other separators between words and sentences. There is also usually no differentiation between upper and lower case.

Another difficulty is created by the absence of a single (standardized) spelling of many words from older periods. In different sources or even within the same document the same word may be found spelled differently, and the lack of dictionaries for the corresponding period makes it extremely difficult to train the recognition tools. In addition to this, it is a well-known fact that the “re-edition” of old books was done by copying them, a human activity that was not free from errors. These errors, once introduced, were proliferated in subsequent copies, and checking against the original was often impossible.

The frequent use of abbreviations, for which there is no standard register, is characteristic of the texts of the period under examination. For those who were in charge of writing the documents, this procedure was a reduction of their efforts, helping them to save space and time. However, the actual deciphering of these abbreviations presents a particular recognition task, which is more difficult to automate.

A special issue is writing over the line. Letters, or combinations of letters, positioned on the line (either to save space or to correct an error) must be recognized separately, and then inserted inside the word in the appropriate position. Alongside these, there are often margin notes, added either during the initial writing process or later.

And since, as mentioned above, human work is not exempt from errors, it was only natural that the author or copyist would want to correct them, often by erasures and writing over the places subject to erasure, which made the writing less clear or destroyed the medium (paper or other).

Despite these challenges, some studies have achieved promising results in recognizing similar texts (including the segmentation of *scripta continua* into words) using neural networks applied to Slavic, Latin, French, and Romanian manuscripts (Rabus, 2019; Clèrice, 2020; Burlacu, 2021).

In this paper we will extrapolate and develop the experience gained from the digitalisation of old Romanian printings, applying it to medieval manuscripts.

In this paper, we explore multiple approaches to address these challenges. We first used ABBYY FineReader versions 15 and 16 to recognize manuscripts written in Romanian Cyrillic uncial script. By developing and training two OCR models tailored for uncial and semi-uncial texts, specifically focusing on 3 manuscripts: *Floarea Darurilor*, *Codicele Bratul*, and *Codex Sturdzanus* — we achieved an accuracy of character recognition of approximately 75-90% (calculated from the ratio of the number of correctly recognized characters to the total number of characters). Additionally, we experimented with AI models like ChatGPT (based on GPT-4) for OCR. Utilizing the OpenAI API and interactive training methods, we assessed the potential of these models in interpreting complex manuscripts. We also discuss implementing an OCR system using a vector database and a pre-trained ResNet50 model for character recognition, which showed promising results, especially when limited training data was available. Finally, we developed a machine learning model based on a DistilBERT architecture to automatically segment texts written without spaces. By simulating *scripta continua* using a corpus from the 1646 *New Testament*, we achieved high accuracy in predicting word boundaries.

2. Related Works

Handwritten Text Recognition initially appeared as a sub-task of Optical Character Recognition, but due to specific requirements and difficulties, today HTR is considered a sibling task (Aguilar, 2024.) HTR and OCR separation was mainly provoked by the problem of digitalisation of ancient manuscripts. The irregular features of handwriting, abbreviations, and scribal idiosyncrasies; variations in spellings; non-standard page layouts; and deteriorated pages - these factors make the application of printing-based OCR difficult. Together with the arrival of AI-based techniques, another factor has been added: HTR of ancient documents is hard to supply by ground truth data. The generating of ground-truth data is always resource- and time-consuming, but the ancient documents HTR also requires employing experts in ancient languages and script domains.

Thus, modern HR research can be summarized as follows. Extending development in layout analysis and creating corpora, researchers mostly concentrated on resolving ground truth data problems (Hawk, 2019.) Following another modern tendency of the OCR domain, HTR researchers propose online (zettacloud_web) and desktop open-source platforms (Reul, 2022.) Finally, DeLORo – the research project that inspired and supported our work has to be mentioned (Cristea et al., 2022).

3. Optical Character Recognition of Medieval Manuscripts

This section describes the materials and methodologies used for recognizing three manuscripts written in Romanian Cyrillic uncial script. To address the challenges of recognizing these manuscripts, we first used ABBYY FineReader versions 15 and 16. This software allowed us to develop and train OCR models specifically for uncial and semi-uncial texts.

3.1. Development of Specialized Models

In this subsection, we discuss the development of specialized models tailored for recognizing Romanian uncial and semi-uncial texts, focusing on refining OCR accuracy through custom datasets and advanced preprocessing techniques.

The common problem of ancient documents digitalisation is the absence of resources for OCR training. Moreover, printed ancient documents can be grouped into several classes, for example, according printing house. But handwritten documents require unique training for every “hand”. We resolve resources lack problem by the usual approach: creating datasets from already digitised documents. Being by far the most common, this approach, however, is time-consuming and prone to human error. To correct the issues, we propose the OCR results processing module.

The choice of format of dataset items is defined by the requirements of modern digitalisation. Initially, digitalisation was considered as the transfer from only readable scanned document images to digital text suitable for search and copy. Today the result of digitalisation is the exact digital copy of the document. Such copy must support fonts, styles, images, and layouts. According to these requirements, docx format is chosen for training dataset items. The final digital copy must be ready for sharing and presentation, so the results of digitalisation will be published in *pdf*.

The errors identification remains a manual job until there is enough referencing data to employ AI. However, the manual correction, that reviewer has to do, consists only in the marking the founded error and the supplying the suggestion of correct text. As the OCR results are presented in *docx* format, the most effective marking is changing of document style parameters. The example is presented in Figure 1. In actual documents, marks are done by colours. In Figure 1, for better black-and-white visibility, marks are done by changing styles: bold for errors and italic for suggestions.

кҀ шпхеровїим їилл(м)ї єйь чєжкҀ єдѣрєлл(м)олпєь

Figure 1: Text example

The transliteration of this excerpt (after correcting all errors, not just the marked ones) is as follows: "... [fo]cu și heruvimii miei ceia cu vedeare multă."

The extraction of the text with a particular value of style parameter is well implemented by docx processing libraries, in our module python-docx. Extracted errors are classified: can they be corrected automatically or must be corrected manually? In the example from Figure 1 the constant error is demonstrated. In this particular handwriting the letter "m" is recognized as "лл". After the writing rules of the considered century were checked, the double letter usage wasn't found. Thus, this error was classified as an error that can be corrected by simple replacements, and the corresponding module was executed. The corrected text can be used in training and corrected again and corrected again while the error frequency is not under the considered threshold. After that, the document will be added to the dataset.

3.2. Processing the Manuscript *Floarea Darurilor*

Floarea Darurilor is a well-known book that was widely circulated in Romanian culture. Originally published in 1491 in Florence, Italy, under the title *Fiore di Virtù*, its Romanian translation was distributed in manuscript form until 1701.

The text is written in uncial script with superscripts, lacking spaces between words and sentences, and without paragraph divisions. It is bilingual, presenting parallel texts in Church Slavonic and Romanian. An example of an original page is shown in Figure 2.

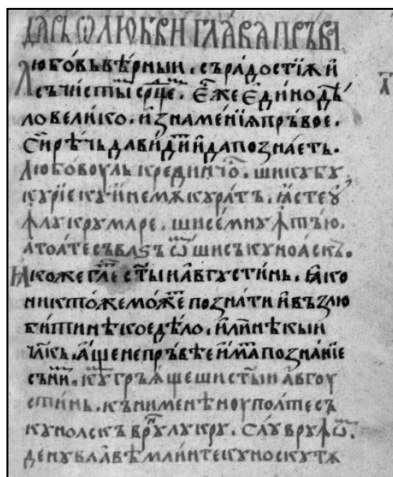
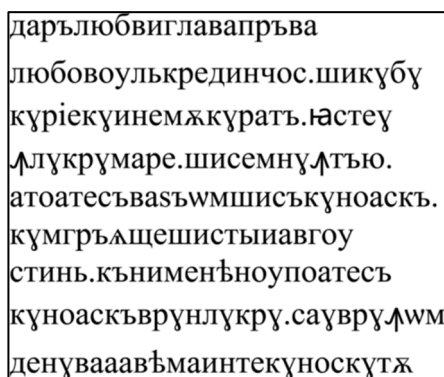


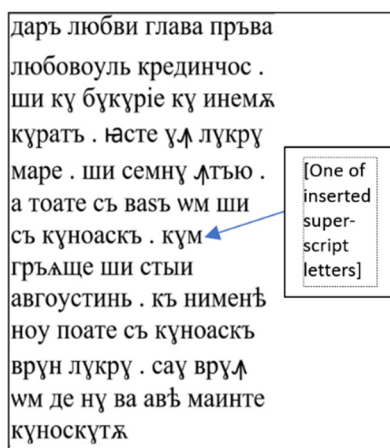
Figure 2: A page from the manuscript *Floarea darurilor*



дарѣлюбвиглавапрѣва
любовоулькрединчос.шикубѹ
кѹріекѹинемжкѹратѣ.ѣстеу
ѿлѹкрѹмаре.шисемнѹѿтѹю.
атоатесѣваѣѡмшиѣсѣкѹноаскѣ.
кѹмгрѣѡщеѣистѣиавгоу
стинѣ.кѣнименѣноупоатесѣ
кѹноаскѣврѹнлѹкрѹ.саѹврѹѿѡм
денѹвааавѣмаинтекѹноскѹтѣ

Figure 5: Corrected recognized text without spaces

Word division was also done manually, and superscript letters were added by inserting them into the text (see Figure 6). Automating this process requires a specialized and more complex solution. The training set comprised approximately 2,000 characters, while the test set included around 300 characters. The recognition accuracy without word division is 85%.



дарѣ любви глава прѣва
любовоуль крединчос .
ши кѹ бѹкѹріе кѹ инемж
кѹратѣ . ѣсте ѿ ѿлѹкрѹ
маре . ши семнѹ ѿтѹю .
а тоате сѣ ваѣѡм ши
сѣ кѹноаскѣ . кѹм
грѣѡще ши стѣи
авгоустинѣ . кѣ нименѣ
ноу поате сѣ кѹноаскѣ
врѹн лѹкрѹ . саѹ врѹѿ
ѡм де нѹ ва авѣ маинте
кѹноскѹтѣ

[One of inserted superscript letters]

Figure 6: Corrected recognized text with inserted spaces.

3.3. Processing the Manuscript *Codex Sturdzanus*

The second source we worked with is *Codex Sturdzanus* – a collection of ancient texts that were copied between 1580 and 1619 by the priest Grigore from Măhaci. Text recognition was conducted using FineReader v.15. The character-level accuracy exceeds 85% after training on a 20-page model (Bumbu, 2024.) The recognition results are displayed in Figure 7.

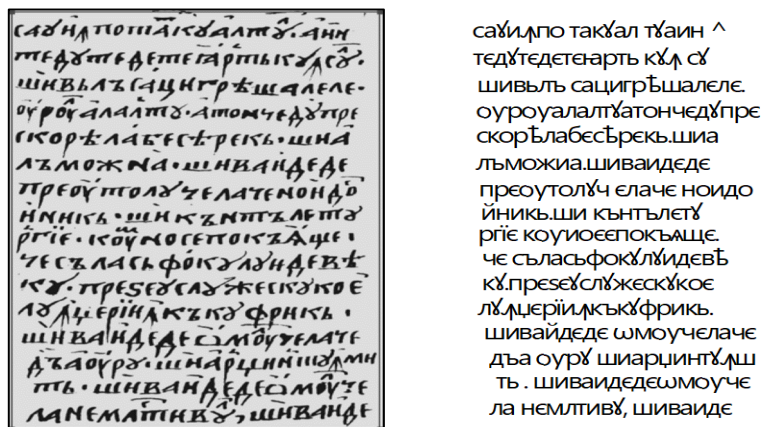


Figure 7: Page from *Codex Sturdzanus* and recognized text

The performance evaluation of a specific OCR model for the *Codex Sturdzanus* indicated an accuracy of approximately 85.6%.

3.4. Processing the Manuscript *Codicele Bratul*

Codicele Bratul is a miscellaneous manuscript from the 16th century that contains several religious texts interleaved with Slavonic and Roman languages. Similar to *Codex Sturdzanus*, the Slavonic sentences are written in black, while the Roman ones are in red.

Text recognition was conducted using FineReader PDF 15. For this purpose, a dataset containing pages from *Codicele Bratul* was prepared. The training set included approximately 4,877 glyphs, while the test set comprised 495 characters. The evaluation focused on the accuracy of OCR at the character level. As a result, 45 characters were misrecognized, resulting in an accuracy rate of 90.91% for FineReader 15.

4. Applications of AI in OCR

The integration of AI in OCR, especially with large language models like ChatGPT, has significantly progressed. When recognizing medieval manuscripts in uncial script, ChatGPT (based on GPT-4) was compared to traditional OCR software like ABBYY FineReader. Unlike traditional models requiring extensive training datasets, ChatGPT leverages pre-trained knowledge for character recognition with minimal extra training, which is beneficial for historical texts

with limited data. Evaluations using the OpenAI API and ChatGPT interface demonstrated its potential in interpreting complex and degraded texts, such as the *Codicele Bratul*.

4.1. Using OpenAI API

Using the OpenAI API for OCR tasks involving historical scripts, such as the uncial manuscript style, offered benefits and drawbacks. The API's main advantage is its ability to handle large volumes of text. Since the API allows for rapid requests, it is ideal for batch-processing large datasets without requiring continuous manual input. This is because the API can be integrated into automated scripts or applications that programmatically send multiple OCR requests, giving the possibility to scale the processing of documents. This cannot be done from the current ChatGPT user interface.

4.1.1. Optimized Prompt Method

The stateless nature of the API presents challenges, as each call operates independently without memory of previous interactions. This means the API does not keep any information between requests. It does not remember anything from earlier calls. Therefore, each time we make a request, we should include all the character details in the prompt. To create a prompt, we sent the model images of each character and it helped us to generate a description, assuming it could later use these descriptions to correctly identify characters during further API calls.

```
{
  "user_text": "Now I will show you the letter \"\u00d0\u00b3\" which looks like the russian letter \"\u00d0\u00b3\" (Unicode: U+0433).",
  "response_text": "The letter \"\u00d0\u00b3\" generally resembles the Russian letter \"\u00d0\u00b3,\" which looks similar to an inverted \"L.\"- The structure consists of a vertical stroke on the left side and a horizontal line extending to the right from the top of the vertical stroke.- In some variations, the horizontal line might have a slight upward or downward slant."
}
```

Figure 8: Optimized prompt with character description using OPENAI API

Figure 8 shows an image containing the response text which is a generated prompt from a previous call used to describe each character. This method achieved 65% accuracy, because of its “stateless nature”, which does not retain all the information of the context between requests. Therefore, this is a negative feature.

4.1.2. Sending Multiple Character Images Method

Another approach involving the API was to send the model images containing handwritten letters from the uncial script, each image containing about 15 examples of a hand-written character. This gave the model a variety of samples to

learn the small differences in each character, helping it recognizes different handwriting styles more accurately.

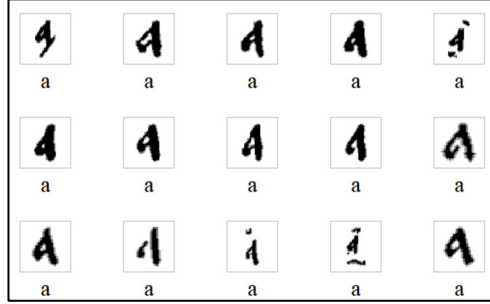


Figure 9: Training image example

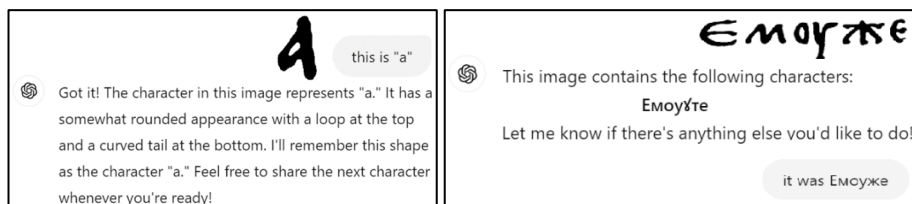
Figure 9 shows an example of an image used in this method; it shows multiple instances of the handwritten character 'a'. Using this approach enhanced the model's consistency in recognizing the uncial script. However, there is a limit of 20 images that can be sent per request, which is not enough to send an image for each letter in the alphabet. Another drawback is the high cost of sending multiple images, making the approach less cost-effective. The method achieved only 50% accuracy due to the limited number of characters that could be shown to the model.

4.2. Using the ChatGPT Interface

In contrast to the API approach, the ChatGPT interface offers a more interactive training process, which is particularly advantageous for iterative OCR tasks involving complex and varied scripts. We employed two distinct approaches to train the model using the ChatGPT interface.

4.2.1. Training with Feedback

In this approach, we initially showed the model of each character from the alphabet one at a time. Then, we tested how well the model remembered each character's appearance by presenting images of individual letters and asking it to identify the letter shown. We tested each letter individually, and if the model made a mistake, we provided feedback by telling it the correct letter. We continued this process of showing an image and providing feedback until the model no longer made any mistakes. We then gradually increased the complexity by showing images of two characters at a time, continuing to train the model with feedback. Once the model reliably recognized two letters, we increased the number of letters to three, four, five, and six, each time providing feedback to correct the model as needed.

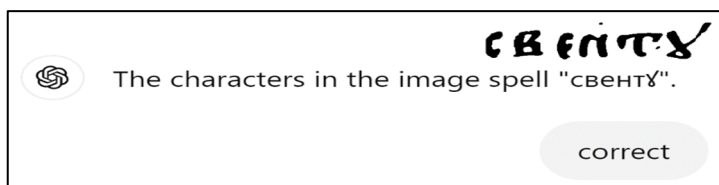


(a) Single Character Stage

(b) Six Characters Stage

Figure 10: Training with feedback

Figure 10 depicts the steps of training, beginning with individual character recognition (a) and progressing to multi-character groups (b). This iterative feedback loop enabled instantaneous modifications, resulting in improved accuracy over time (Figure 11). It assisted the model in learning and recognising specific character shapes and stylistic nuances of the uncial script more efficiently.

**Figure 11:** Result example for a word with 6 letters

The ChatGPT interface is a versatile environment for character recognition, enabling efficient recognition for individual characters and short text segments, which proved advantageous for historical texts with diverse character shapes. However, recognizing longer text required progressively more iterations — 20 for two-character images and up to 40 for five-character ones — making it labor-intensive and impractical for large-scale projects. The model achieved 90% accuracy for words up to six letters, thanks to detailed feedback, but its accuracy dropped for longer texts due to complexity, limiting scalability.

4.3. Using Vector Database

We are currently building a new OCR system using a Weaviate vector database and a pre-trained ResNet50 model for character recognition. The system starts by segmenting the input image into individual letters using image processing techniques like contour detection. Each segmented character is vectorized through ResNet50, and these vectors are stored in the Weaviate database to facilitate efficient similarity searches. Recognition is then performed using k-nearest

neighbours (k-NN), which compares the vector of the new character with the stored ones.

We discovered that the system can achieve impressive performance in recognizing characters, even when provided with just one reference image per letter. At this stage, the system is capable of processing images containing only a single character. During tests on a dataset of 500 characters, the system reached an accuracy of 93% using 10 reference images per letter. We are actively working on improving the system's performance and scalability, particularly in addressing challenges like character segmentation for historical texts with connected or overlapping letters.

4.4. The Accuracy of the Discussed Methods

The accuracy of each method shown in Table 1 was evaluated using the same page containing 500 characters.

Table 1: Accuracy results

Method	Accuracy	Notes
Optimized prompt method	65%	Accuracy could be improved with a better system prompt.
Handwritten characters image method	50%	The accuracy was impacted due to the limitations on the number of pages that could be sent.
Training with feedback method	75%	Accuracy decreases significantly for words longer than 6 letters, making it impractical for longer sequences.
Vector database-based OCR system	93%	Used only 10 reference images per character.
The best result for ABBYY FineReader model (<i>Codicele Bratul</i>)	90%	Trained on 4877 characters.

5. Segmenting *Scripta Continua* Texts

Segmenting *scripta continua* texts – manuscripts written without spaces between words – is a significant challenge in historical document analysis. Our study aims to develop a machine learning model to automatically segment such texts, facilitating transcription and interpretation of 16th-century manuscripts. The lack of a sufficiently large dataset created based on texts from the period of interest is the biggest impediment to solving this problem. To overcome this, we proposed to use the 17th-century corpus at our disposal (being aware of the difference in writing and spelling) to make easier the expansion of the small dictionaries we have at our disposal.

For dataset preparation, we utilized a corpus from the 1646 *New Testament* written in Cyrillic script, sourced from Bumbu (2020) [1]. Although this corpus did not feature *scripta continua*, we simulated the format by removing spaces, converting the text into a continuous stream of characters like some 16th-century manuscripts.

Example of text from corpus [1]:

Ши дѣка пости патрѹѹечи де зиле, ши патрѹѹеч де нопци, апои флѣмжнѹи.

Transformed text:

Шидѣкапостипатрѹѹечидезиле,шипатрѹѹечденопци,апоифлѣмжнѹи.

At model training we used a DistilBERT-based neural network adapted for token classification, labelling each character as the beginning of a word ('B') or inside a word ('I'). Training used 4,000 samples, with 1,000 samples reserved for testing. The model learned to predict word boundaries by recognizing language patterns and contextual cues in continuous character sequences. Initially, the model struggled with unique special characters. To address this, we augmented the tokenizer with these characters and retrained the model to handle symbols such as 'Ѣ', 'Ѧ', 'Ѩ', 'Ѫ', 'Ѭ', 'Ѯ', 'Ѱ', 'Ѳ', 'Ѵ', 'Ѷ' etc. On the test set, the model achieved a precision of 1.0000, recall of 0.9977, and F1 score of 0.9989, indicating effective prediction of word boundaries in *scripta continua* text. Some samples of the results are shown in Figure 12.

Unsegmented text:	Ширѣспѹнѣрѣ,нѹщимѣдеѹнде.
Segmented text:	Ширѣспѹнѣрѣ, нѹ щимѣ де ѹнде.
Unsegmented text:	ЛѣсѣѹдеашимѣрѣсѣрѣѦгалилеа.
Segmented text:	Лѣсѣ ѹдеа ши мѣрѣсѣ ѣрѣ Ѧгалилеа.
Unsegmented text:	Ельзисѣ:адѹчецимѣеачѣлѣѦкоачѣ.
Segmented text:	Ель зисѣ: адѹчеци мѣе ачѣлѣ Ѧкоачѣ.
Unsegmented text:	Апроповѣдѹианѹлдомиѹлѹипрѣимить.
Segmented text:	Апроповѣдѹи анѹл домиѹлѹи прѣимить.

Figure 12: Example output from the model applied to test data

We applied the model to authentic 16th-century manuscripts which might have much-unsegmented text (*scripta continua*). Although we tested only a few examples, preliminary results are promising. The model effectively segmented the

text in many cases (see Table 2), but further testing is needed to make definitive conclusions.

Table 2: Examples from *Codex Sturdzanus*

Unsegmented text	Segmented text	Comments
чєпѡвєци	чє пѡ авєци	—
плѣнгѡдопѡвой	плѣнгѡ допѡвой	Should be "плѣнгѡ допѡ вой". In the 17th century, "допѡ" is written as "дѡпж" (după).
плѣнгѡдѡпжвой	плѣнгѡ дѡпж вой	Experimenting with the 17th-century spelling "дѡпж".
чалачєнаѡлѣѡсачѣла	чала чє наѡ лѣѡс ачѣла	—
милѡципрєсєѡ	милѡ ци прє сєѡ	Should be "милѡци прє сєѡ". Both "милѡ" and "милѡци" are valid words, creating ambiguity.
фокѡдєвѣкъ	фокѡ дє вѣкъ	—
вапостикѡкѡръцїєачєлѡ	ва пости кѡ кѡръцїє ачєлѡ	—
скаѡноулмїєѡчєласфнтѡл.	скаѡн оул мїєѡ чєла сфнтѡл.	Should be "скаѡноул мїєѡ чєла сфнтѡл." Both "скаѡн" and "скаѡноул" are valid words, creating ambiguity.

This study demonstrates the potential of machine learning for segmenting the words in *scripta continua* manuscripts. While the model performed well on some data we have now, applying it to more manuscripts needs to be done. Preliminary results are promising, but more extensive testing is required to confirm the model's effectiveness.

6. Conclusions

Digitising medieval Romanian manuscripts written in Cyrillic uncial script poses significant challenges due to varied handwriting styles, inconsistent spelling,

copying errors, complex edits, extensive abbreviations, and continuous writing without spaces (*scripta continua*). We addressed these issues by developing specialized OCR models using ABBYY FineReader, achieving 75-90% accuracy on texts like *Floarea Darurilor*, *Codicele Bratului*, and *Codex Sturdzanus*. While we experimented with AI models like ChatGPT GPT-4, ABBYY FineReader outperformed them in our OCR tasks. Additionally, our DistilBERT-based model effectively segmented *scripta continua* texts with high accuracy. Combining traditional OCR with modern AI techniques enhances the digitalisation and accessibility of medieval.

Acknowledgments

This work was supported by project 011301 “Information Systems based on Artificial Intelligence”. Substantial support in the creation of the dataset was given to us by colleagues from the Institute of Theoretical Informatics (IIT), the branch of the Romanian Academy, Iași, who developed the DeLoRo project in partnership with the Romanian Academy Library (Bucharest) to whom we hereby thank.

References

- Aguilar, S. T., Jolivet, V. 2023. Handwritten Text Recognition for Documentary Medieval Manuscripts. *Journal of Data Mining and Digital Humanities*, 2023, Historical Documents and Automatic Text Recognition, {10.46298/jdmdh.10484}. {hal-03892163v3}
- Bumbu, T., Burtseva, L., Cojocaru, S., Colesnicov, A. and Malahov, L. 2023. Digitalisation of Romanian Historical Printings. “Vladimir Andrunachievici” Institute of Mathematics and Computer Science, Moldova State University. Chișinău, 2023 (Valinex), 176 p. ISBN 978-9975-68-497-2
- Bumbu, T. 2020. On Alignment of Textual Elements in a Parallel Diachronic Corpus. *Computer Science Journal of Moldova*, Volume 28, Issue 3, page 384.
- Bumbu, T., Burtseva, L., Cojocaru, S., Colesnicov, A., Malahov, L. 2024. Medieval Text Processing: Use Case of Romanian Uncial Writing. *Proceedings of the International Conference dedicated to the 60th anniversary of the foundation of “Vladimir Andrunachievici” Institute of Mathematics and Computer Science*.
- Burlacu, C., Rabus, A. 2021. Digitising (Romanian) Cyrillic using Transkribus: new perspectives. *Diacronia*, 14, December 12.
- Clèrice, T. 2020. Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin. <https://jdmdh.episciences.org/6264/pdf>
- Cristea, D., Rebeja, P., Pădurariu, C., Onofrei, M., Scutelnicu, A. 2022. Data Structure and Acquisition in DeLoRo - a Technology for Deciphering Old Cyrillic-Romanian Documents. *Proceedings of ConsILR 2022*.
- Hawk, BW, Karaisl, A and White, N 2019. Modelling Medieval Hands: Practical OCR for Caroline Minuscule, *Digital Humanities Quarterly*, vol. 13, no. 1.

- Rabus, A. 2019. Recognizing handwritten text in Slavic manuscripts: A neural-network approach using Transkribus. <https://www.academia.edu/38835297/>
- Reul, C., Tomasek, S., Langhanki, F. and Springmann, U., 2022, *Open source handwritten text recognition on medieval manuscripts using mixed models and document-specific finetuning. International Workshop on Document Analysis Systems* (p. 414-428). Cham: Springer International Publishing.

NEW ACHIEVEMENTS IN THE DIGITISATION OF LINGUISTIC ATLASES OF THE ROMANIAN LANGUAGE

SILVIU-IOAN BEJINARIU¹, LUDMILA MALAHOV², TUDOR BUMBU²,
FLORIN-TEODOR OLARIU³, RAMONA LUCA¹

¹ *Institute of Computer Science, Romanian Academy Iași Branch,
2, T. Codrescu Street, Iasi, Romania*
{silviu.bejinariu,
ramona.luca}@iit.academiaromana-is.ro

² *“Vladimir Andrunachievici” Institute of Mathematics and Computer Science,
Moldova State University, 5, Academiei Street, Chisinau, Republic of Moldova*
{ludmila.malahov, tudor.bumbu}@math.usm.md

³ *“Alexandru Philippide” Institute of Romanian Philology, Romanian Academy
Iași Branch, 2, T. Codrescu Street, Iasi, Romania*
olariuft@yahoo.com

Abstract

The paper presents the latest functionalities that enrich the information system developed for the digitisation of dialect information and the publication of linguistic atlases. The most important new function is dedicated to the digitisation and republication of the Moldovan Linguistic Atlas using the Latin script. Given that the dialectal information still exists only in printed format with the Cyrillic alphabet, a procedure for optical recognition and transliteration of answers on the map is proposed in order to enter them into the dialectal database. The system is also adapted for the automatic generation of maps in the Moldovan Linguistic Atlas. Another new research tool useful for the dialectologists is the automatic generation of reports with the indexes of words and forms. The indexes are generated as Word documents which makes them available not only for consulting but also for their usage in larger studies. A new graphic representation is also proposed. It allows the automatic creation of interpretive maps with the frequency of appearance of the forms.

Keywords: linguistic atlas, phonetic transcription, optical character recognition, geographic information system.

1. Introduction

The development of information systems for publishing linguistic atlases of the Romanian language began more than two decades ago in the academic centres of Iași and Cluj-Napoca and still continues with the creation of new tools useful for dialectologists. The *AlrMaps* application (and its previous versions) is one of the information systems developed by the researchers in the Institute of Computer Science (IIT), Romanian Academy Iași Branch in collaboration with dialectologists from the “Iorgu Iordan – Al. Rosetti” Institute of Linguistics (ILIR) and “A. Philippide” Institute of Romanian Philology (IFRI) since the year 2000. It was used to publish the volumes III, IV and V of the *New Romanian Linguistic Atlas by Regions – Moldavia and Bukovina* (NALR-MB), the second volume of the *Linguistic Atlas of the Aromanian Dialect* (ALAR) and the *Linguistic Atlas of Romanian Dialects from the North and South of the Danube* (ALDRo) which is in press. Also, other information systems developed within the same collaborations have been used to publish the *Atlas Linguarum Europae* (ALE) vol. I, fasc. IX, the *Romanian Linguistic Atlas by Regions – Wallachia and Dobrogea* (ALRR-MD) vol. VI and now are used to prepare the *Romanian Linguistic Atlas by Regions. Synthesis* (ALRR.Synthesis) vol. IV.

2. The Moldovan Linguistic Atlas

The study of the Romanian language spoken east of the Prut River has been a continuous concern of dialectologists in the Republic of Moldova. In 1956, the research on Moldovan dialects of the Romanian language across the entire territory of the former Soviet Union began, with the major objective being the creation of a linguistic atlas that would include material comparable to that of the linguistic atlases published up to that time (Popovschi et al., 2022).

Between 1957 and 1965, dialectal material was collected from 240 localities in Bessarabia, Transnistria, northern Maramureș (Transcarpathia), northern Bukovina, Hertsa region, South-Eastern Ukraine, Caucasus (Abkhazia and Krasnodar Krai in the Russian Federation), and the Asian part of the former USSR (Kazakhstan, Kyrgyzstan, and Primorye in the Russian Far East) as it is shown in Fig. 1.

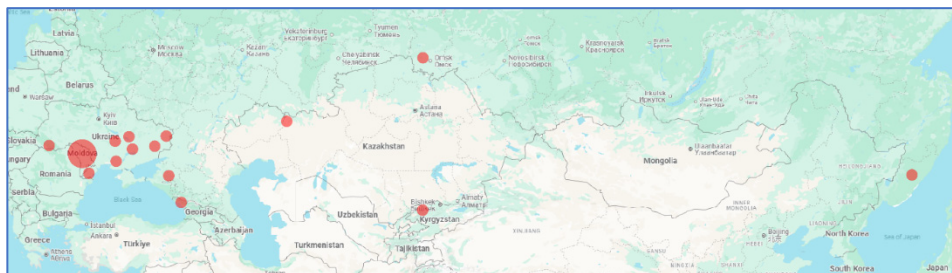


Figure 1: Distribution of the ALM survey points across the former Soviet Union

2.1. The Atlases Already Published Using the Cyrillic Alphabet

The dialectal material that contains the answers obtained from the questions in the predetermined questionnaire were used to create eight volumes of the linguistic atlas. Four volumes (actually two volumes, each of them with two parts) of *The Moldovan Linguistic Atlas* (original title in Romanian: *Atlasul lingvistic moldovenesc* – ALM) were published before 1990 using the Cyrillic alphabet (Fig. 2). After 1990, other four volumes were published, this time using the Latin script and with a new title: *The Romanian Linguistic Atlas by Regions. Bessarabia, Northern Bukovina, Transnistria* (original title in Romanian: *Atlasul lingvistic român pe regiuni. Basarabia, nordul Bucovinei, Transnistria* – ALRR–Bas).



Figure 2: The cover and a map from ALM vol 1, part 1

About half of the texts recorded on magnetic tapes during dialectal surveys were published between 1969–1987 in 6 volumes of the dialectal texts. Supplement to the *Moldovan Linguistic Atlas* (original title in Romanian: *Texte dialectale. Supliment la Atlasul Lingvistic Moldovenesc*). These texts are a valuable addition to the data from the linguistic atlases, because they allow a more efficient understanding of the mechanisms of speech which is essential for spoken language studies.

Given the undeniable value of the dialectal texts published in Cyrillic alphabet before 1990, the “B. Petriceicu-Haşdeu” Institute of Romanian Philology from Chişinău (IFRC) in collaboration with researchers from the “Vladimir

Andrunachievici” Institute of Mathematics and Informatics (IMI) started in 2020 a project for the revitalization of dialectal texts as part of the national cultural heritage. As a result, in 2023 was published the volume *Dialectal Texts. Supplement to the Moldovan Linguistic Atlas / Romanian Linguistic Atlas by Regions. Bessarabia, Northern Bukovina, Transnistria*, vol I, part I. (Fig. 3). For the publication of the volume using Latin script, computer technologies were used for character recognition and transliteration.

2.2. Digitisation of Moldovan Linguistic Atlases

As part of a Romania – Republic of Moldova collaboration project, we propose in this paper a system for the digitisation and republication in Latin script of Moldovan linguistic atlases from before 1990. To achieve this goal, the experience gained by IIT and IFRI in the publication of Romanian linguistic atlases will be used on the one hand, and the experience of the IFRC and IMI in digitising old texts in Cyrillic script on the other hand. However, in contrast to dialectal texts that contain lines of text that are relatively easy to detect, linguistic maps contain several types of information: sequences of 1-2 words representing the answers in phonetic transcription, numbers of the survey points, titles and notes, and for all

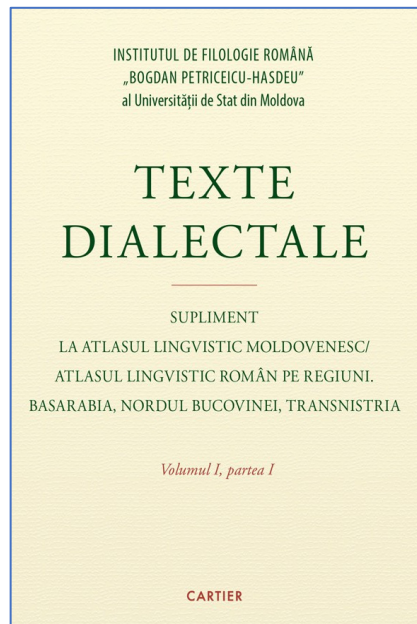


Figure 3: Cover of the Dialectal Texts volume published in 2023

of these the positioning on the map is very important. Thus, the model of the information system is depicted in Fig. 4.

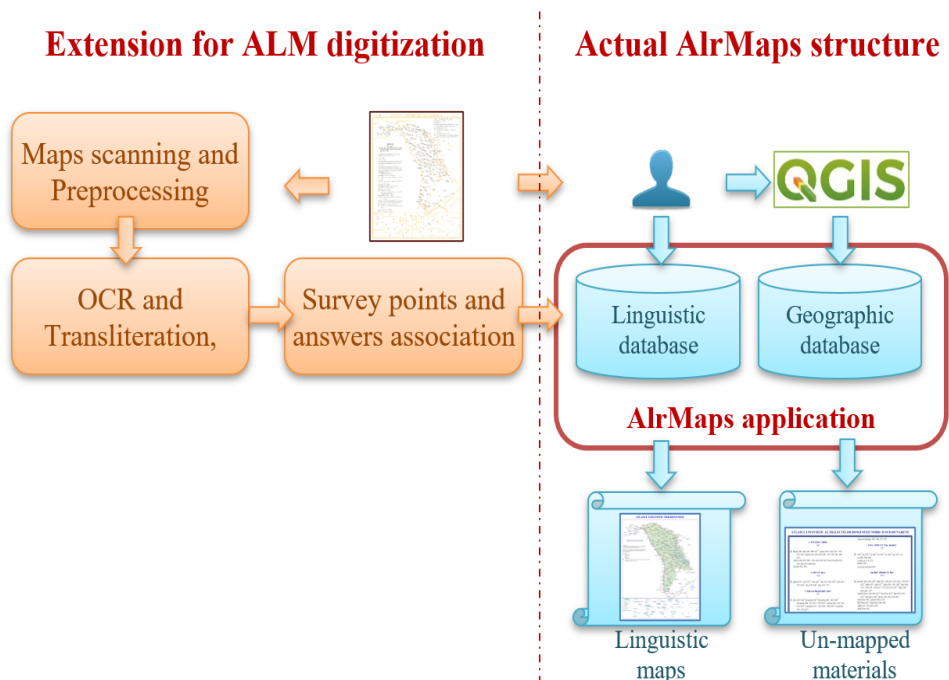


Figure 4: Structure of the information system for ALM digitisation

In the right side of Fig. 4 the structure of the existing *AlrMaps* application (Bejinariu et al., 2022c; 2023a) is depicted. The geographic database content must be prepared using the QGIS package for each linguistic atlas. It contains the map template (geographic features which will be drawn on maps) as well as the page layout (space reserved for titles, notes, images). It is organised on layers, which makes linguistic maps fully customisable. The linguistic database contains the list of survey points and the list of title words. The database also contains, for each word, the phonetic transcription of the answers obtained in each survey point. This is a sketched presentation, in fact each record contains a lot of related information as supplementary position description, notes, comments, images and even the audio recording of the answers (if available). Due to the very large number of possible symbols (~115,000), the multiple positioning and formatting modes at the symbol level, each of them is binary encoded on 10 bytes. Also, the symbols in phonetic transcriptions are drawn in real-time by an image synthesis module of *AlrMaps*. All these make the phonetic transcriptions thus represented usable only in specialized applications (Bejinariu et al., 2022a). As output, the *AlrMaps* application generates the linguistic maps which can be saved in several image file

formats or directly printed. The maps can also be saved in an XML compatible format for later editing. By changing the drawing properties or repositioning some elements on the map, it can be significantly improved from an aesthetic point of view before printing. Several reports that contain the list of words, survey points, transcriptions of the answers as well as indexes of forms can also be generated in Microsoft Word file format (Bejinariu et al., 2022b; 2023a; 2023b).

As it is designed, the *AlrMaps* application currently allows editing the content of the linguistic database of ALM by directly entering the answers by the dialectologist. In fact, the map presented in Fig. 9 was made this way, because the implementation of the processing chain for automatic digitisation is in progress.

The ALM digitisation module is an extension of *AlrMaps* application. As input, the digitisation module uses the image files obtained by scanning the pages of ALM. After applying several preprocessing steps and the optical characters recognition functions, the answers on the map are converted in text files which are then transliterated in the Latin alphabet. Finally, the module has to rebuild, for each map, the associations between the answers and the survey points to which they belong.

The researchers from IMI and IFRC have a vast experience in the digitisation of old writings, proven by the publication of volumes of dialect texts, but also by the digitisation of old texts from the 17th-19th centuries (Bumbu et al., 2022, 2023; Cojocaru et al., 2017a; Cojocaru et al., 2017b; Colesnicov et al., 2018; Popovschi et al., 2022).

The entire digitisation process of the ALM maps is briefly detailed in the following paragraphs.

A. Maps Scanning

The linguistic maps from the ALM volumes published before 1990 using the Cyrillic alphabet still exist only in printed form. For scanning these maps, it must be taken into account that they are approximately 50 years old. Therefore, from the available volumes the maps in the best condition must be chosen for scanning. However, the size of the image files must also be taken into account considering the processing that must be applied. Thus, the decision was to scan the maps with the resolution 600 dpi, full colour (24 bits/pixel). To avoid artifacts and quality loss due to compression, the images are saved in TIFF files using LZW lossless compression and their size varies between 30 and 40 MB. Because the size of the maps is non-standard, somewhere between A3 and A2 (38x46 cm), it was necessary to scan them using a professional equipment.

B. Image Preprocessing

Considering that the maps are almost 50 years old, the quality of the images is not very good, at least in some cases. For this reason, it is necessary to apply a

preprocessing step for automatic page orientation correction, colour correction, words breaking and noise cleaning. For this purpose, the ScanTailor, a free application which perform corrections was used.

C. Text Extraction

There are many cases in which the elements drawn on the linguistic map are overlapped, e.g., answers and borders or survey points numbers; bounding boxes of the answers or even the answers between them (Fig. 5.a). Obviously, these situations lead to wrong OCR results. In order to avoid this or at least to reduce the impact of overlaps, a procedure is proposed for extracting the elements of interest, i.e., the texts that are the subject of the OCR procedure.

As can be seen in Fig. 2, the maps are drawn in two colours. The map's template, which is drawn in orange (or brown, depending on the page quality) is fixed. The page title, page frame, borders, cities, rivers, names of cities and rivers, the numbers of the survey points, boxes and regions names at the bottom are the same on all maps, thus it is not necessary to recognize them. It is obvious that their position relative to the page frame is always the same, but the absolute position in the image may vary depending on the placement of the page when scanning.

Only the elements drawn in black are of interest and are therefore subject of the text recognition process. In this category are: the title word, the question to which it answers, the phonetic transcriptions of the answers as well as the notes placed on the left side of the map.

Text extraction is done by applying two simple image processing methods:

The colour planes (R, G, B) are extracted from the colour obtaining three images in grey levels, one for each component.

The elements of the map's template, initially drawn in orange have a higher intensity (>200) in the red plane while the text to be recognized has lower intensities (< 50). Thus, by applying a simple thresholding on the red plane (e.g. threshold = 128) leads to obtaining an image that contains only the text of interest.

An example is presented in Fig. 5.

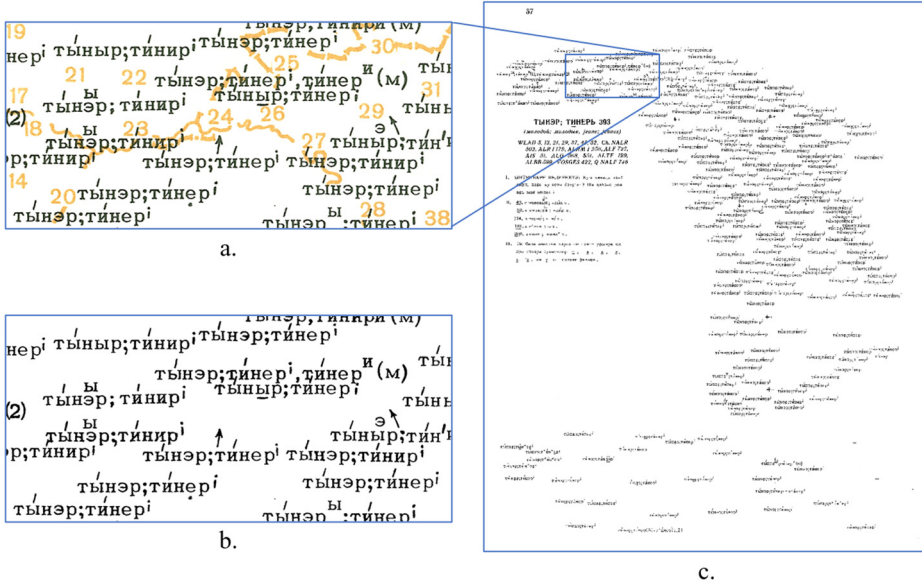


Figure 5: Text extraction result; a. sample from original image; b. extracted text from sample image; c. extracted text from full map

D. Optical Character Recognition and Errors Correction

After the text is extracted and the image is prepared, the optical characters recognition is performed by using the Abbyy FineReader application. The processing is made by using a Romanian language pre-trained-model written in phonetic Cyrillic. The OCR results (Fig. 6) is then checked in order to correct recognition errors due most of the times to the overlapping of the answers on the map.

Since the texts to be recognized represent a transcription of dialect speech with nuanced pronunciations, it is impossible to use a predefined dictionary for the automatic correction of the OCR result. However, as part of the project to digitise the volume of dialect texts, colleagues from the IFRC and IMI have created such a dictionary, which is still being completed. For each answer detected on the map, in addition to the OCR result, its position in the image is also retained. The position is required to restore associations between answers and survey points in subsequent steps.

The OCR step is still being implemented, but its principles are described in detail in (Bumbu et al., 2023; Cojocaru et al., 2017; Popovschi et al., 2022).



Figure 6: Result of the OCR saved in XML format

E. Automatic Transliteration and Errors Correction

Then a transliteration is performed from the phonetic alphabet based on the Cyrillic script to the one based on the Latin script. The Cyrillic alphabet has about 80 letters with and without diacritics, 16 diacritics, as well as ligatures and separators. Thus, after recognition, there are approximately 100 characters for transliteration. The transliteration process uses a set of 273 already defined rules (Popovschi et al., 2022). Some examples of rules and their application are shown in Table 1.

Table 1: Some transliteration rules (Popovschi et al., 2022)

Transliteration rules			Text sample
Cyrillic alphabet		Latin alphabet	
ý	→	ú	пýни → púni (put on...)
ÿ	→	u	брыÿ → briu (belt)
ʝóá	→	ʝóá	сʝóарили → sʝóarili (sun)
уы	→	uâ	доуы → douâ (two)
ʝ + vowel	→	ʝ + vowel	ʝограды → ʝográdâ (courtyard)

F. Establishing the Association of the Answers with the Survey Points

As mentioned above, the result of the digitisation process does not include the numbers of the survey points, but even if they are recognized, the answers must be associated with the points.

The spatial coordinate of the survey point is used to locate nearby phonetic transcriptions. Typically, a transcription is located above its corresponding survey point, but it can sometimes be positioned to the right, left, or below, often indicated by an arrow. These arrows serve as valuable clues for automatic association. In Fig. 7, for the survey points 29, 40, and 47, the corresponding phonetic transcriptions are placed below the points, and the arrows help in directly linking the transcriptions to the correct survey points.

Our method is based on the K nearest neighbours (KNN) technique which can help us identify survey points and the phonetic responses in their proximity. In cases where arrows are not present, a proximity criterion based on a maximum allowable distance is applied to identify neighbouring transcriptions. Additionally, most phonetic transcriptions start with the same letter as the nearby responses. By verifying if there is sufficient space before the initial letter to exclude other characters, we can correctly associate the transcription with the survey point. Once the first letter of a phonetic transcription is identified, the entire line of transcription can be extracted. Figure 8 illustrates the survey point 71, where transcriptions are positioned very close to each other. In this case, point 71 can be associated with two possible responses. We save both associations, allowing an expert to disambiguate and select the correct one.



Figure 7: Phonetic words indicating survey points with arrows

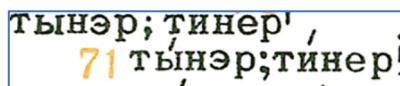


Figure 8: Survey point with closely positioned responses

In situations where the same transcription could be associated with multiple survey points, we retain all possible associations for further validation as we research the points. The validation stage is important to ensure data accuracy, especially in densely populated areas on the map.

Some results obtained by applying steps A – F for two maps in ALM are shown in Tables 2 and 3. In Table 2 is also figured a wrong result of the OCR. For point 30, the word “СКЭП” is recognized as “СКЭН”.

Table 2: Results obtained for three points in map 10 – I/you let go
(‘am/ai dat drumul’ in Romanian)

Point	Transcription image	OCR result	Transliteration result
30	СКЭП; СКЭК"	СКЭН; СКЭК"	scăn/ scăk
31	СКЭП;СКЭТ"	СКЭП; СКЭТ"	scăp/ scăt'
32	СКЭП; СКЭТ", СКЭК" (Ф)	СКЭП; СКЭТ", СКЭК"(Ф)	scăp/ scăt'/ scăk(f)

Table 3: Results obtained for three points in map 21– Saddle (‘Şa’ in Romanian)

Point	Transcription image	OCR result	Transliteration result
30		ша́ ỳы	şá ûâ
31		ша́ ỳы	şá ûâ
32	ШЭ	ШЭ	şă

G. Importing Data into the *Alrmaps* Dictionary

In the final step, the transliterated answers have to be imported in the *AlrMaps* dictionary. Due to the different way of encoding the information (especially the phonetic phenomena) a data conversion function is necessary which will be implemented when the previous stages are fully functional. Considering the small number of symbols and phonetic phenomena used in ALM, the data conversion will be done with the help of a translation table.

2.3. ALM Maps Preparation

The geographic database which contains the map template and page layout are prepared using the free open-source QGIS package as it was already described in (Bejinariu et al., 2023 a, 2023b). All map elements are organized in layers which makes it fully customizable. By using a simple interface in which the user has only to select the title word from the linguistic database, the new maps are instantly generated (Fig. 9).

Since the proposed system is still being implemented, the answers on the map in Fig. 9 were not obtained by the processing described above. In this case, the transliterated answers were entered into the linguistic database by dialectologists using the editing interface of the *AlrMaps* application.

Unlike the other atlases created using the *AlrMaps* application, in ALM there are some maps with a very large number of notes that complete the answers. Since these exceed the space reserved on the page on which the map is also located, they are presented on a separate page, similar to those with un-cartographed materials. Therefore, in the case of ALM, the *AlrMaps* application will generate two different pages if needed.

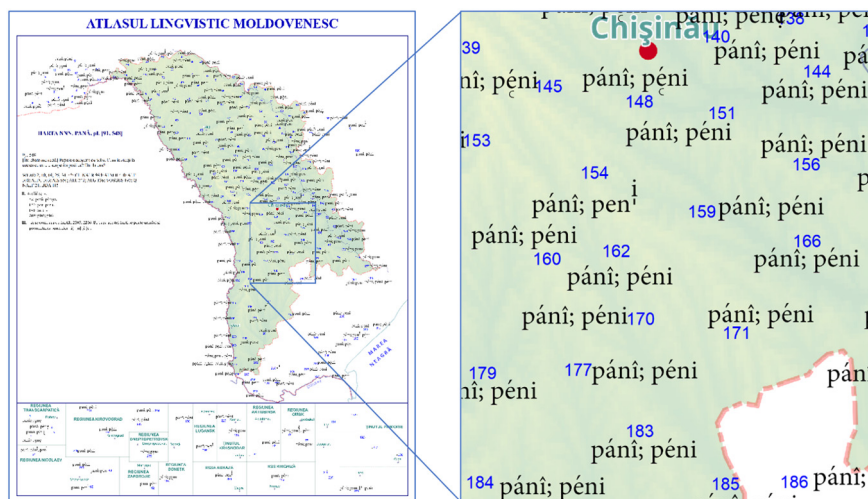


Figure 9: Full map and detail from the automatically generated ALM drawing with Latin script

3. Tools for Indexes of Words and Forms Generation

Often, in their research, dialectologists evaluate the number or frequency of certain words or phonetic forms in the responses obtained in during surveys. It is obvious that the storage of information in a database allows these evaluations to be carried out instantly. In the following paragraphs, two such tools are presented that differ in the way of delivering the results, as text documents and graphic images respectively.

3.1. Interface for Automatic Generation of Indexes

The interface of the *AlrMaps* application allows the automatic generation of indexes by using the “Index of forms” command on the ribbon bar. When selected, *AlrMaps* shows the dialog box in Fig. 10. First of all, the user must edit the search form using the phonetic transcriptions editing interface. Then, the search mode must be defined using the following options: whether or not to search in the phonetic transcriptions of the answer, note and comment, to search as phonetic transcription and/or as transliterated text, to perform case sensitive search, whether to ignore or not the legato symbols. The user has the possibility to add two filters: search for all or for a certain title word and search for all or for a single survey point.

Index de cuvinte și forme

☒ Filtru cuvânt: (albina) ÎNȚEAPĂ (ind. prez. 3)

☐ Filtru punct anchetă:

Forma de căutat: ae

Caută în:

- ☒ Transcriere
- ☐ Nota II
- ☐ Comentariu

Mod căutare:

- ☒ Text
- ☒ Transcriere fonetică
- ☐ Ignoră legato
- ☐ Case senzitiv

Mai multe opțiuni...

Salvează indexul în: D:\Silviu\Projects\ALR_New\Data\Dictionary\MB_Demo_MN\new_index_list.docx

Caută! Renunță

Figure 10: Interface for indexes automatic generation

More detailed search options are available if the command “More options” is selected. The application shows the “Search form description” dialog box in Fig. 11. The search form is displayed, but it is not editable, only the characters in the form can be selected one at a time. The first option allows selecting the position of the searched form in the word: prefix, suffix or any position. The other selections refer to the characters in the form. The vertical position (normal, overlapped or half-raised) can be selected as well as the groups of phenomena sought. Depending on the current selection, the groups of phenomena specific to vowels and consonants are displayed. For each of these, the application or not of the filter related to the group of phenomena can be indicated, and respectively if a phenomenon from the respective group is applied or not.

After completing the search criteria selection, the report is generated automatically. The search result is saved in tabular format in a Microsoft Word document whose name must also be indicated by the user.

Descrierea formei de căutat...

Forma căutată:

Selectje:

Poziționare în răspuns: ☐ Prefix ☐ Sufix ☒ Oricare

Poziționare verticală:

Filtru selectat:

Forma căutată: e
Poziția în răspuns: Oricare

1: 'e' (selectat)
Poziție: normal
vocală

Vocală - fenomene aplicate

☐ Lungime:

☐ Nazalizare:

☐ Coup de glotte:

☐ Deschidere:

☐ Afonizare:

☐ Semivocală:

Figure 11: Advanced search options for indexes automatic generation

3.2. Synthetic Map with Frequency of Occurrence of Forms

Another tool of interest for dialectologists is the possibility to generate synthetic maps that show the frequency of forms occurrences. The interface is the same as when generating the indexes, with the difference that at the end, the default hatch mode of the map must be set according to the number of occurrences determined. The application shows the “Occurrence frequency” dialog box presented in Fig. 12. The form which is the subject of the map, the minimum and maximum number of occurrences is displayed in non-editable fields.

Frecvența de ocurență....

Secvența căutată: e

Număr minim de apariții: 11

Număr maxim de apariții: 33

Număr intervale: 5

Limita minimă: 10

Limita maximă: 35

Culoare minim: [white swatch]

Culoare maxim: [blue swatch]

OK Cancel

Figure 12: Selection of the default drawing mode for the occurrence's frequencies

The system proposes a way of dividing the number of occurrences into equal intervals. Thus, for a number of 5 intervals, the lower limit is the largest multiple of 5 that is less than or equal to the minimum number of occurrences and the upper limit is the smallest multiple of 5 that is greater than or equal to the maximum number of occurrences. Also, the system proposes the colours associated to the first and last interval of occurrences. For the intermediate intervals the colour varies evenly between these two colours. The user has the possibility to change all these values. Because the synthetic maps are always related to a way of grouping the survey points for a title word, in case of the frequencies map, the system creates a grouping mode for a virtual word entitled “_dummy”.

So far, the frequency map is completely defined and can be generated. However, the user can customize the default drawing mode using the group editing interface (Fig. 13) which is detailed in (Bejinariu et al., 2022b). The final result is presented in the Fig. 14 that shows a map made for Linguistic Atlas of Romanian Dialects from the North and South of the Danube - ALDRo.

Cuvânt: _dummy_

Titlu legendă: ocurențele formei: e

Forme cuvânt (6):

Nr	Legendă	Puncte asociate
1	nr. ocurențe: 31+35	012, 025, 044
2	nr. ocurențe: 26+30	008, 020, 022, 026, 027, 028, ...
3	nr. ocurențe: 21+25	003, 005, 006, 007, 009, 011, ...
4	nr. ocurențe: 16+20	001, 002, 010, 014, 015, 016, ...
5	nr. ocurențe: 11+15	004, 009, 008, 009, 009, 009, 010, ...

Puncte asociate formei curente (55):

Punct	Nume	Comentariu
003	Voivozi	
005	Doba	
006	Valea Vinului	
007	Negrești-Oaş	
008	Casa	

Legenda formei selectate

nr. ocurențe: 21 ÷ 25

Stil contur:

Culoare contur:

Stil interior:

Culoare interior:

Simbol:

Culoare simbol:

Puncte (0):

Punct	Nume	Transcriere

Figure 13: Detailed drawing settings for the frequencies map

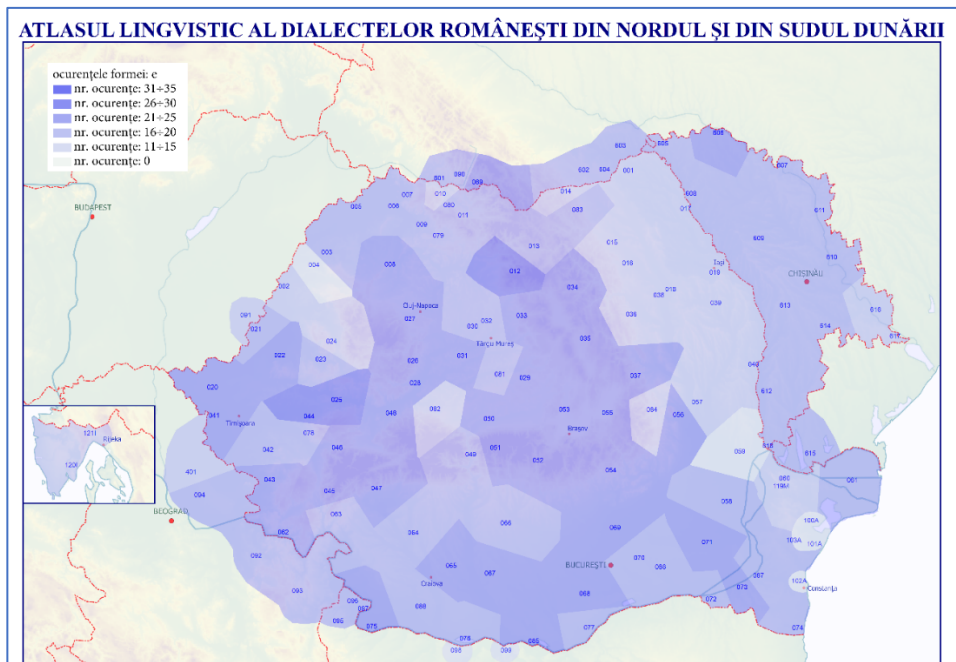


Figure 14: Example of frequency map and its legend

4. Conclusions

In this paper, the new achievements in the digitalisation of linguistic atlases are proposed. The main contribution is the system for the republication of a new edition of the Moldovan Linguistic Atlas using the Latin script. Offering the possibility of studying the phonetic forms in the entire space of the Romanian language, the atlas that will be published will represent an important addition to

European linguistics and to Romanian and Romance dialectology in particular. As for the tool for indexing words and forms, it helps researchers in conducting extensive studies on the variation of Romanian dialects over time and space.

Acknowledgements

The work presented in section 2 was supported by a grant of the Ministry of Research, Innovation and Digitisation, CNCS–UEFISCDI, project number PN-IV-P8-8.3-ROMD-2023-0318, within PNCDI IV. The results presented in section 3 of this paper were obtained within the framework of the fundamental research projects of the Romanian Academy.

References

- Silviu-Ioan Bejinariu, Ramona Luca, Vasile Apopei and Florin Iftene. 2022a. Editarea textelor dialectale folosind transcrierea fonetică specifică limbii române (EN: *Dialect text editing using the phonetic transcription of the Romanian language*). In *Anuar de lingvistică și istorie literară* (EN: *Yearbook of Linguistics and Literary History*), Tome LXII. In Memoriam of Stelian Dumitrăcel, pages 37-53.
- Silviu-Ioan Bejinariu, V. Apopei, M. Nevaci, F. Iftene, C.I. Floarea and N. Saramandu. 2022b. Realizarea hărților interpretative pentru Atlasul Lingvistic al Dialectelor Românești din Nordul și din Sudul Dunării (EN: *Design of interpretive maps for the Linguistic Atlas of Romanian Dialects from the North and South of the Danube*). In *Fonetică și Dialectologie* (EN: *Phonetics and Dialectology*), XLI, Bucharest, pages. 1-11.
- Silviu-Ioan Bejinariu, R. Luca, F. Iftene, M. Nevaci, C.I. Floarea and N. Saramandu. 2022c. GIS Based Publishing System for the Romanian Linguistic Atlas by Regions – Synthesis. In Svetlana Cojocaru, Victoria Bobicev, Tatiana Verlan, Dan Tufiş and Dan Cristea (editors), *17th edition of The International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2022*, 10-12 November 2022, Chişinău and online, pages 81-92.
- Silviu-Ioan Bejinariu, Vasile Apopei, Manuela Nevaci, Florin-Teodor Olariu and Nicolae Saramandu. 2023a. Information Technology and Geolinguistics. In *12th Conference on Speech Technology and Human-Computer Dialogue – SpeD 2023*, October 25-27, 2023, Bucharest, Romania.
- Silviu-Ioan Bejinariu, Florin Iftene, Manuela Nevaci and Nicolae Saramandu. 2023b. Information Technology based Research Tools for Dialectology, *Xth Congress of the International Society for Dialectology and Geolinguistics*, September 4-8, 2023, Bucharest, Romania.
- Tudor Bumbu, Lyudmila Burtseva, Svetlana Cojocaru, Alexandru Colesnicov and Ludmila Malahov. 2022. A platform for Processing Heterogeneous Documents. In Svetlana Cojocaru, Victoria Bobicev, Tatiana Verlan, Dan Tufiş and Dan Cristea (editors), *17th edition of The International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2022*, 10-12 November 2022, Chişinău and online, pages 81-92.
- Tudor Bumbu, Lyudmila Burtseva, Svetlana Cojocaru, Alexandru Colesnicov and Ludmila Malahov. 2023. Distinctive features of recognition for documents printed in

- the Romanian transitional alphabets. In *Computer Science Journal of Moldova*, vol.31, no.3(93), pages 340-350.
- Svetlana Cojocaru, Constantin Ciubotaru, Alexandru Colesnicov, Ludmila Malahov and Tudor Bumbu. 2017a. Instrumentar pentru digitalizarea și transliterarea textelor tipărite în limba română cu caractere chirilice (EN: *Tool for the Digitisation and Transliteration of Printed Texts in Romanian with Cyrillic characters*). In *Revista Bibliotecii Academiei Române* (EN: *Journal of the Library of the Romanian Academy*), Year 2, Nr. 2, Jul.-Dec. 2017, pages 27-38.
- Svetlana Cojocaru, Alexandru Colesnicov, Ludmila Malahov. 2017b. Digitisation of Old Romanian Texts Printed in the Cyrillic Script. *Proc of the Second International Conference on Digital Access to Textual Cultural Heritage DATECH2017*. Goettingen, June 1-2, 2017, pages 143-148.
- Alexandru Colesnicov, Svetlana Cojocaru, Ludmila Malahov and Lyudmila Burtseva. 2018. On Text Processing after OCR. In Vasile Păiș, Daniela Gîfu, Diana Trandabăț, Dan Cristea and Dan Tufiș (editors), *Proceedings of the 13th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Iași, 22-23 Nov. 2018, pages 53-60.
- Liliana Popovschi, Ludmila Malahov and Vlada Colesnicova. 2022. Digital processing of dialectal texts published in Chisinau in the years 1969–1987. In *AKADEMOS* Nr. 4(63)/2021, pages 48-53.

Linguistic Atlases

- Atlas Linguarum Europae. Vol. I. fasc. IX, European Linguistic Maps (Original title in French: *Cartes Linguistiques Européennes*). 2015. N. Saramandu, M. Nevaci, I. Geană (coords). *Bucharest University Press*. Available at: <https://lingv.ro/atlas-linguarum-europae/>.
- Texte Dialectale. Supliment la Atlasul Lingvistic Moldovenesc / Atlasul Lingvistic Român pe Regiuni. Basarabia, Nordul Bucovinei, Transnistria (EN: *Dialectal Texts. Supplement to the Moldovan Linguistic Atlas / Romanian Linguistic Atlas by Regions. Bessarabia, Northern Bukovina, Transnistria*). 2023. Vasile Pavel (scientific editor); Liliana Popovschi (editor); Ludmila Malahov (digital processing). Chișinău: 2023, *Cartier*, Vol. 1, Part 1, 189 pages. ISBN 978-9975-86-772-1.
- Atlasul lingvistic al dialectului aromân (EN: *The Linguistic Atlas of the Aromanian Dialect*) vol. II. 2020. Nicolae Saramandu and Manuela Nevaci. Bucharest, Publishing House of the Romanian Academy. Available at: <https://lingv.ro/wp-content/uploads/2021/11/Atlasul-lingvistic-al-dialectului-aroman-ALDA.pdf>.
- Atlasul Lingvistic Moldovenesc (EN: *The Moldovan Linguistic Atlas*). 1968. vol. I: Rubin Udler, part 1, Fonetica (EN: *Phonetics*); Rubin Udler, Vasile Melnic, part 2, Fonetica, Morfologia (EN: *Phonetics, Morphology*), Chișinău, 1968, maps 1-520; 1972. vol. II: Victor Comarnițchi, part 1, Lexicul (EN: *The lexicon*), Chișinău, 1972, maps 521-721; 1973. Vasile Melnic and Vasile Pavel, part 2, Lexicul (EN: *The lexicon*), Chișinău, 1973, maps 722-1034.
- Noul atlas lingvistic român pe regiuni. Moldova și Bucovina (EN: *The New Romanian Linguistic Atlas by Regions. Moldova and Bukovina*) vol. V. 2022. Materials collected by Vasile Arvinte, Stelian Dumistrăcel, Ion Florea, Ion Nuță, Adrian Turculeț and edited by Stelian Dumistrăcel, Luminița Botoșineanu, Florin-Teodor Olariu, Veronica Olariu,

Alexandru-Laurențiu Cohal, Iași, Publishing House of “Alexandru Ioan Cuza” University of Iași.

Atlasul lingvistic român pe regiuni. Muntenia și Dobrogea (EN: *The Romanian Linguistic Atlas by Regions. Wallachia and Dobrogea*). vol. VI. 2020. By Teofil Teaha, Nicolae Saramandu and Bogdan Marinescu. Edited by Nicolae Saramandu and Manuela Nevaci. Bucharest, Publishing House of the Romanian Academy.

CHAPTER 2

CORPORA

ROMEMES: A MULTIMODAL MEME CORPUS FOR THE ROMANIAN LANGUAGE

VASILE PĂIȘ, SARA NIȚĂ, ALEXANDRU-IULIUS JERPELEA,
LUCA PANĂ, ERIC CUREA

Research Institute for Artificial Intelligence "Mihai Drăgănescu",

Romanian Academy, Bucharest, Romania

vasile@racai.ro

saramaria.nita9@gmail.com

alex.jerpelea@gmail.com

lucapana571@gmail.com

Abstract

Memes are becoming increasingly more popular in online media, especially in social networks. They usually combine graphical representations (images, drawings, animations or video) with text to convey powerful messages. In order to extract, process and understand the messages, AI applications need to employ multimodal algorithms. In this paper, we introduce a curated dataset of real memes in the Romanian language, with multiple annotation levels. Baseline algorithms were employed to demonstrate the usability of the dataset. Results indicate that further research is needed to improve the processing capabilities of AI tools when faced with Internet memes.¹

1. Introduction

Memes represent visual content obtained by combining one or multiple images with text. In some cases, instead of static images, the meme creators use animations or videos, while in some other cases the image is completely missing, being replaced by a simple background colour. Furthermore, sometimes the text is missing, the message being conveyed only by visual elements. Regardless of their form, memes can be found in all types of online media, particularly in social networks. They are usually humorous in nature and are used to convey powerful messages that are spread rapidly by Internet users (Bauckhage, 2011). Memes are usually inspired by events happening at the time of their creation. Thus, they can be related to sports events, politics, culture, etc. For example, in a period

¹ This work is licensed under a Creative Commons Attribution 4.0 International Licence. License details: <http://creativecommons.org/licenses/by/4.0/>.

preceding election, we expect to see memes related to political parties or candidates. The message can be either positive or negative with regard to the represented subject. Hristova (2014) acknowledges that political imaginaries of the structure and scope of the collective will are intimately tied to iconic images that shore up, reinforce, and make literal and legitimate political visions. During the 2020 US presidential election, the social web was flooded with user-generated Internet memes commenting on one of the most important political events worldwide. Internet memes have become a widespread format through which social media users can express individual opinions (political or otherwise) (Johann, 2022). AI applications trying to analyse the message expressed in memes must employ multiple algorithms, adapted to their multimodal nature. Thus, image analysis must be combined with natural language processing. Furthermore, certain memes contain a combination of images, requiring additional segmentation. Results obtained from individual analysis processes must be fused together to obtain the final result.

This paper introduces a new corpus of memes in the Romanian language, called RoMemes². It contains multiple levels of annotations, making it suitable for multiple tasks, such as: text recognition, simple sentiment analysis, complex sentiment analysis, fake image detection, identification of political memes. Our contribution is twofold: first, we create and release the curated corpus of Romanian memes, and second, we perform experiments to demonstrate the usability of the corpus for research purposes and establish baseline results. The rest of the paper is organized as follows: Section 2 presents related work, Section 3 presents the methodology used, Section 4 describes the corpus, Section 5 gives experimental results, and we conclude in Section 6.

2. Related Work

Vyalla and Udandara (2020) acknowledge the increased interest in both meme generation and propagation. They propose a framework for automatically generating memes and introduce a dataset consisting of 1.1 million meme captions from 128 classes. Kiela et al. (2020) focus on hate speech detection in multimodal memes. They acknowledge that seemingly harmless texts and images when combined can become hateful. Thus, unimodal analysis is not enough and multimodal algorithms must be employed. Hwang and Shwartz (2023) introduce the MemeCap dataset, containing 6,384 memes, for the task of meme captioning. For this purpose, an AI model is presented with a meme and the associated post title (assuming the meme is extracted from social media) and the model must

² <https://zenodo.org/doi/10.5281/zenodo.13120215>

generate a relevant caption, describing the meaning of the meme. The authors acknowledge the challenging nature of the proposed task, due to the metaphorical nature of the meme content. Tanaka et al. (2022) classified a set of 7,500 memes as being humorous or not. Sharma et al. (2020) describe a SemEval challenge consisting of three subtasks: sentiment (positive, negative, and neutral) analysis of memes, overall emotion (humour, sarcasm, offensive, and motivational) classification of memes, and classifying intensity of meme emotion. The best performances achieved were F1 (macro average) scores of 0.35, 0.51 and 0.32, respectively for each of the three subtasks. For the challenge, 1,000 meme samples were provided as trial data, 6,992 samples as training data while 1,879 samples were used as test data. Sharma et al. (2024) introduced MOOD (Meme emOticOns Dataset), which embodies six basic emotions, and proposed a neural framework named ALFRED (mechanism emotion-Aware muLtimodal Fusion foR Emotion Detection) for meme emotion detection.

Many datasets, including the one proposed in this paper, suffer from class imbalance. To mitigate this problem, Chawla et al. (2002) proposed SMOTE, a synthetic minority over-sampling technique. This approach uses both the under-sampling of the majority class and a particular kind of over-sampling of the minority class by creating “synthetic” examples. Moreover, frameworks for data augmentation, such as NL-Augmenter (Dhole et al., 2023), have been proposed, allowing automatic enrichment of less represented classes.

Considering the Romanian language, different resources and tools are useful for analysing the text component of a meme. The Romanian Wordnet (Tufiş et al., 2004) is a large lexical database of nouns, verbs, adjectives, and adverbs, grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. The Romanian Emotion Lexicon (RoEmoLex v.3) (Lupea and Briciu, 2019) includes associations between a set of words and two sentiment orientations (positivity and negativity) as well as eight basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust). It was originally translated from an English version, the NRC Word-Emotion Association Lexicon (EmoLex) (Mohammad and Turney, 2010), but now has extra tags, new words and phrases, and synonyms for the original terms. Ciobotaru and Dinu (2021) created the RED dataset for emotion recognition in Romanian tweets. CoRoLa (Tufiş et al., 2019) is the Reference Corpus for Contemporary Romanian Language. It has been used for creating word embeddings used for text representation (Păiş and Tufiş, 2018). Avram et al. (2022) proposed a distilled BERT (Devlin et al., 2019) model for the Romanian language. Given the short text usually associated with memes, social media datasets and models may be relevant when analysing memes. In this context, Tăiatu et al. (2023) proposed the RoBERTweet model together with a system that detects emotions more accurately than earlier general-domain Romanian and

multilingual language models by utilising the RoBERTweet model. Păiș et al. (2022a, 2022b) introduce the MicroBloggingNERo, a dataset of Romanian short messages, specific to micro-blogging platforms.

3. Methodology

The corpus of Romanian memes was created at the Research Institute for Artificial Intelligence “Mihai Drăgănescu” of the Romanian Academy. The activity involved 5 annotators working under the supervision of a senior researcher. The annotators team contained a combination of researchers and young students. Before the start of the activity, an “Annotation guidelines” document was created and shared with the annotators. Furthermore, a series of discussions (online and via email) were used to familiarize the annotators with the task. Throughout the annotation process, communication with the annotators team was maintained through online means. All ambiguities were discussed and the annotation guidelines were updated when needed.

The annotators received access to a personal shared Google Drive folder. Each annotator was responsible for collecting their own memes from social media or through Google searches. They were instructed to select only memes with Romanian language text. Following the gathering of new memes, a Google sheet file was filled with information corresponding to each meme. We gathered the following data:

- file name – typically this is the file name generated by the platform hosting the meme. However, in some cases this was renamed by the annotator (for example when the platform always produced files with the same name, such as “image.jpg”). In the case of renaming, the original extension was kept. No conversion took place on the files;
- source – the source from where the meme was downloaded. This is not necessarily the first place where the meme was uploaded. As memes tend to become viral, being spread extensively by Internet users, it is difficult to determine the actual source of the meme;
- URL – the Internet address associated with the page hosting the meme (the location from where it was downloaded);
- text – this is the text from the meme. It was extracted exactly as it appears in the meme, including special characters where these were present. We focused only on the message of the meme, while additional text elements (such as logos, messages appearing on clothing) were not extracted, if they were considered not relevant for the meme;
- complexity – this field refers to the complexity of the image. We distinguish between: simple images and combinations of multiple images;

- **real/fake** – indicates the authenticity of the image(s) used for the meme. We distinguish between 3 possible values: **Real** – the image seems to be a real image repurposed for the meme; **Fake** – the image is a clear fake (such as elements taken from other images superimposed on a base image or containing drawings); **DeepFake** – the image is fake, but this is hard to tell without additional world knowledge (it looks like a real image, but the annotator knows it cannot be true);
- **sentiment1** – this field is used for simple sentiment analysis, including 3 classes: **Positive** - the meme conveys a positive sentiment (generally speaking, or towards a person or group); **Negative** – the meme conveys a negative sentiment; **Neutral**: the meme conveys no clear positive or negative sentiment;
- **sentiment2** – this field is used for complex sentiment analysis. We consider the 6 emotion categories described by Parrot (2001): joy, love, fear, anger, surprise, sadness. Memes in general are supposed to be amusing. However, the annotation takes into account the emotion that is conveyed by the actual content (situation, person, animal, etc.) of the meme;
- **political** – identifies political memes (related to political parties or individuals involved in politics).

Each annotator had an initial target of 100 memes. We kept only memes that did not contain specific copyright messages. At the end of the annotation process, 50 memes were randomly selected (10 from each annotator) and distributed to be re-annotated by other team members. This allowed us to compute inter-annotator agreement for the different annotation types. The result is given in Section 4.

4. The RoMemes Dataset

Following the collection process, the corpus was processed to eliminate any duplicates or memes containing copyright messages. Furthermore, the files were renamed using a numerical sequence, while keeping the original extension, converted to lowercase. This step was needed because some file names contained very long names or special characters, which may not be usable in all file systems. The dataset comprises three folders and one file: “images” – folder containing image files; “text” – folder containing corresponding text files (each file has the same name as its associated image file, with the .txt extension); “conllup” – folder containing CoNLL-U Plus files with the text tokenized and annotated; “metadata.tsv” – tab separated file, containing metadata and annotations. This file contains the following columns:

- *ID* – the numerical sequence associated with the file (file names are constructed based on this ID and the extension, indicated in the following column for images, or .txt for text). Example: “0030023”;
- *Extension* – the original file extension (jpg, jpeg, png, webp);

- *Complexity* – the complexity annotation (the annotations were described in Section 3, above);
- *Real_Fake* – the real/fake annotation;
- *Sentiment1* – simple sentiment annotation;
- *Sentiment2* – complex sentiment annotation, based on 6 emotions;
- *Political* – annotation with “Yes”/”No” values indicating if the meme covers political aspects;
- *Width* – width of the image;
- *Height* – height of the image;
- *Channels* – colour channels of the image;
- *Mime* – the mime type associated with the file. This is not always the same as can be deduced from the extension. This indicates a possible problem with the platform hosting the memes, from which the corresponding file was downloaded. For example, 2 files with jpg extension actually have a mime type of “image/png”, 1 file with jpg extension has mime type of “image/webp”;
- *ImageFileSize* – is the size in bytes of the image file;
- *TextSizeChar* – is the size in characters of the associated text file;
- *TextSizeBytes* – is the size in bytes of the associated text file. The text file is saved in UTF-8 encoding. A difference between the TextSizeChar and TextSizeBytes indicates the presence of multibyte UTF-8 characters (usually meaning the presence of Romanian language diacritic characters, such as “ă”, “î”);
- *TextUppercase* – contains the value “Yes” if the text contains only uppercase characters or symbols;
- *TextLowercase* – contains the value “Yes” if the text contains only lowercase characters or symbols. If both TextUppercase and TextLowercase contain the value “Yes”, then the text is composed only of special characters.

The text part of the corpus was processed in the RELATE platform (Păiș et al., 2020; Păiș, 2020; Păiș et al., 2019) and was automatically segmented, tokenized, and annotated with part of speech tags, lemmas, and dependency parsing information. Basic language processing, within the RELATE platform, was performed using UDPipe (Straka et al., 2016), with a new model (Păiș et al., 2021a) trained on the RoRefTrees (Barbu Mititelu et al., 2018) corpus, version 2.7. This choice is justified by the relative processing speed and accuracy of the

results, as indicated in Păiș et al. (2021a). The resulting files were exported in CoNLL-U Plus format³.

Table 1: Statistics for images and categories of the RoMemes corpus

Indicator	Value	Indicator	Value
Number of memes	462	Polarity Negative	221
Min Width	259	Polarity Positive	89
Max Width	3,839	Polarity Neutral	152
Min Height	194	Emotion Anger	79
Max Height	3,166	Emotion Fear	38
Complexity Simple	358	Emotion Joy	87
Complexity Combination	104	Emotion Love	14
Real Images	293	Emotion Sadness	106
Fake Images	153	Emotion Surprise	138
Deep Fake Images	16	Political	147

Table 2: Statistics for the text part of the RoMemes corpus

Indicator	Value	Indicator	Value
Number of tokens	5,366	UPOS Noun	1,172
Avg. tokens/meme	11.62	UPOS PropN	619
Uniqiue tokens	2,424	UPOS Verb	658
Hapax legomena	1,831	UPOS Adj	157
Sentences	1,074	UPOS Adv	302
Avg. tokens/sentence	5	UPOS Pron	421
Unique lemmas	1,840	UPOS Num	141

³ <https://universaldependencies.org/ext-format.html>, accessed: 31st August 2024.

Statistics for the RoMemes dataset are provided in Table 1, for the images and annotations, and in Table 2, for the text part. Images vary in size from very small (194 pixels) to very large (3,839 pixels). For many of the annotation categories, the corpus is balanced. However, a few of the categories have a small number of samples (DeepFake images – 16, Emotion Love – 14). Thus, the corpus may not be suitable for experiments with these particular categories. The meme’s text is characterized by very short text, having in average 2.32 sentences/meme, and short sentences (an average of 5 tokens/sentence). The text has a lot of nouns (1,791 considering both UPOS Noun and PropN annotations) and there are “sentences” without verbs (658 verbs compared to 1,074 sentences).

5. Experiments

Given the different levels of annotation of the dataset, we performed several experiments in order to set baseline values. These are described in the following subsections, with regard to text extraction, sentiment analysis (including polarity and emotion detection), political memes detection and fake image detection.

5.1. Text Extraction

Text recognition is an essential first-step when developing AI tools that aim to understand memes, proving to significantly increase performance on comprehension tasks (Priyadarshini and Cotton, 2022). Text in memes tends to be partially structured, usually adhering to certain templates (Schmidt et al., 2020), but the background image on which the caption is placed still poses difficulties for current open Optical Character Recognition (OCR) technologies. To evaluate the accuracy of the OCR outputs, we employ two widely-used metrics: BLEU (Papineni et al., 2002) and ChrF++ (Popovic, 2017) (with the standard word order of 2). Both of these metrics indicate the proportion of similar parts from the automatically and manually extracted meme texts. ChrF++ considers both word and character n-grams, as opposed to BLEU which is calculated as a precision only over n-grams. The OCR engine that we use is Tesseract V5 from Google. Tesseract is a technology based on deep learning models with high accuracy and support for many languages (Zacharias, 2020). We experiment with both tessdata⁴ (default) and tessdata-best⁵ (highest accuracy, but slowest) LSTM models for the Romanian language. However, initial assumptions are made by Tesseract OCR for

⁴ <https://github.com/tesseract-ocr/tessdata>, accessed: 31st August 2024.

⁵ https://github.com/tesseract-ocr/tessdata_best, accessed: 31st August 2024.

text recognition, performing best in certain conditions⁶, so preprocessing is needed for our use case. Moreover, the page segmentation mode (PSM) has a significant impact on the quality of the output. After careful finetuning of this parameter, we chose PSM 11 (“Sparse text. Find as much text as possible in no particular order.” (Iskandarani, 2015)).

Preprocessing is performed by first upscaling the resolution of all memes which have a resolution lower than the 300 Dots per Inch (DPI) standard quality. Then, the alpha channel is removed. Bilateral filtering is used to reduce noise. Because its weight has an additional component of difference in density (Dias and Lopes, 2023), this filter preserves edges instead of blurring them. We used a neighbourhood radius of 5, and a spatial sigma of 2 for our filter. We add a small black 10 by 10 pixels border to account for text at the edge of the image. RGB images are converted to grayscale. We applied binarization with a threshold value of 0.9 on a scale from 0 to 1 (where 0 is equivalent to the black colour and 1 to white). Memes are finally deskewed with a threshold of 40%. All the image processing was realized using ImageMagick⁷. Example images, before and after the preprocessing pipeline, are given in Figure 1.

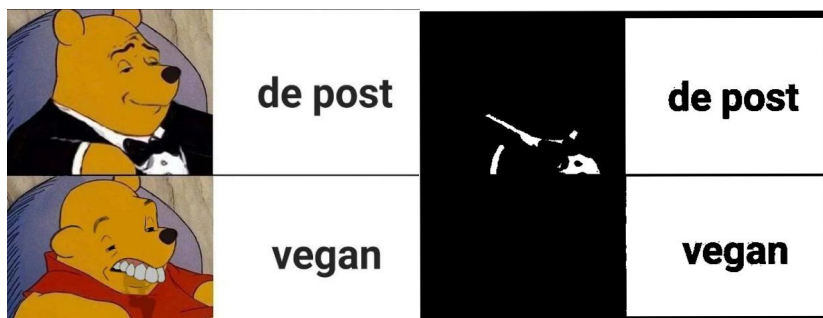


Figure 1: A meme about vegan food, before and after our preprocessing

By manually evaluating samples, we observe that most OCR errors come from either: a) lines with random characters in between the lines that actually contain the text of the meme (Table 3), b) misspelling of words caused by failure of OCR to correctly identify some characters. We also note the efficiency of using machine learning models to correct noisy text from OCR (Nguyen et al., 2021), and aim to correct the outputs from our best scoring methodology with two separate experiments: in the first one, we completely remove some errors, and in the second

⁶ Tesseract documentation. <https://tesseract-ocr.github.io/tessdoc/ImproveQuality>, accessed: 31st August 2024.

⁷ <https://imagemagick.org>, accessed: 31st August 2024.

experiment, we aim to modify the text. We process each line of text, having at most 10 characters (the other lines are kept in order), with a fastText-based language identification model (Joulin et al., 2017). We extract the two languages with the highest probability and remove the line if Romanian is not one of them. The 10 characters limit for the line length was determined through grid search, wanting to achieve the highest BLEU and ChrF++ scores. We tried processing the remaining lines of text through a sequence-to-sequence transformer model, finetuned for Grammatical Error Correction (GEC) of Romanian text⁸. However, this last step reduced the overall accuracy of the text extraction pipeline. We believe this is due to how certain memes are written: without diacritics, using all caps, with punctuation problems, or grammatically incorrect. Table 3 shows the results from the different text extraction experiments.

Table 3: Text extraction results

Model	BLEU	ChrF++
tessdata + upscale	26.74	63.73
tessdata-best + upscale	28.52	64.60
tessdata + full preprocessing	37.35	65.27
tessdata-best + full preprocessing	38.06	65.43
tessdata-best + preprocess + remove lines	45.40	65.67
tessdata-best + preprocess + remove lines + GEC	13.78	55.23

5.2. Sentiment Analysis

We performed sentiment analysis, with regard to both polarity and emotion classification, by employing only the text part of the dataset. We employed the method described in Niță and Păiș (2024). The text was first translated into English using the GoogleTranslator from the deep_translator library to allow compatibility with the NRC emotion lexicon. Given the relatively small dataset, we opted to translate the sentences rather than to translate the lexicon. The text was lemmatised using the WordNet (Fellbaum, 2005) lemmatizer in the NLTK package to help in feature development. Three binary features were created based on the number of words found in the lexicon, indicating a majority of words being positive, negative or neutral in meaning. The same process was utilized in

⁸ <https://huggingface.co/readerbench/RoGEC-mt0-base>, accessed: 31st August 2024.

generating the 8 lexicon-predicted emotion features. If an emotion was the most encountered according to the lexicon, it would be recorded as ‘1’ (when a tie occurred, all of the most-encountered emotions were marked as ‘1’). One aspect to note is that the NRC does not contain data related to how words express ‘love’. Thus, the value of occurrence for this emotion was interpreted as a mean between ‘trust’ and ‘joy’. This choice was motivated by Robert Plutchik’s general psychoevolutionary theory of emotion (1980), in which, beyond the eight primary emotions: Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust, all other emotions are mixtures, or compounds of these primary emotions, including love, which is a combination of joy and trust. A confidence value was also computed based on the number of words belonging to different classes which were found in the lexicon for a given sample.

Additional features were generated using a BERT model for sentiment and emotion predictions on the translated data, trained by Niță and Păiș (2024). The LLM used for training the system was bert-base-uncased. This was chosen due to the fact that the base and large models produced almost the same results (both F1 scores of around 0.3 for polarity and around 0.16 for emotion classification), and thus the smaller system was chosen. The LLM classifier has two additional linear layers, with 2,048 and 1,024 cells respectively, employing ReLU and tanh activation functions. These are followed by a final class prediction head. The model was trained for at least 5 epochs and a maximum of 20 epochs, with early stopping when there was no improvement for 3 epochs. During the first 3 epochs, the LLM was frozen, and only the last linear layers were trained. A batch size of 6 was used. The learning rates for the LLM and the other layers were kept separate. The best hyperparameters were found to be an encoder learning rate of $1.0e-05$ and a linear layers learning rate of $3.0e-05$. The final model training lasted 5 epochs.

The feature set (dataset features as described in Section 4, lexicon features, and BERT features) was split into training, development, and testing datasets and used in a Random Forest algorithm. To tackle the imbalance of the dataset, new examples were synthesized using a SMOTE pipeline (Chawla et al., 2002), and a Stratified K-fold with $n_folds = 5$ was used to avoid overfitting. The best hyperparameters were found using GridSearch. Best hyperparameters for the Random Forest used for polarity: `max_depth: None`, `min_samples_leaf: 2`, `min_samples_split: 2`, `n_estimators: 200`. Best hyperparameters for emotion classification: `max_depth: None`, `min_samples_leaf: 1`, `min_samples_split: 4`, `n_estimators: 100`. Feature selection was then done, removing some of the features in order to achieve a higher F1 score for the test set, while also making sure that the model is still generalizing well. For both polarity and emotion detection, the BERT predictions for negative polarity and surprise were dropped, along with the

feature relating to whether an image is deepfake. Also, the text length was dropped only for emotion classification. For polarity the best results actually occurred when the BERT features were not used. Detailed results are shown in Table 4.

Table 4: Sentiment analysis (polarity and emotion detection) results

Model	Polarity				Emotion			
	Acc	P	R	F1	Acc	P	R	F1
BERT	0.489	0.163	0.333	0.321	0.255	0.420	0.166	0.103
RF without BERT	0.443	0.443	0.443	0.442	0.314	0.323	0.317	0.317
RF with BERT	0.412	0.426	0.414	0.415	0.343	0.346	0.342	0.342

5.3. Political Memes Detection

Memes have a particular appeal to world-wide cultures and are being used by political actors to communicate various political messages with the aim of changing certain beliefs (Beskow et al., 2020). We aim to establish a baseline for detecting Romanian language political memes, in which we only make use of the meme’s text content, encouraging future multimodal input approaches. Although most Large Language Models (LLMs) are not up to date with the political world stage, depending on their knowledge cut-off, they have proven to be useful in various zero-shot learning classification problems (Chae and Davidson, 2023). We experiment with two instruction LLMs: Llama3.1-8B-Instruct (Dubey et al., 2024) and RoLlama3-8B-Instruct (Masala et al., 2024). We feed the LLM each meme text input and prompt it to classify it as political not. The instruction we used has been carefully tuned and stimulates chain-of-thought reasoning (Wei et al., 2022): *Ești un asistent util care clasifică conținutul text al unui ‘meme’ drept politic sau nu. Îți explici pe scurt decizia, iar apoi răspunzi la final cu ‘DA’ sau ‘NU’* (English translation: *You are a useful assistant that classifies the textual content of a ‘meme’ as political or not. You briefly explain your decision, and then answer at the end with ‘YES’ or ‘NO’*). The results are given in Table 5. The low

performance seems to suggest the need for future more complex meme comprehension algorithms, employing multimodal data.

Table 5: Political classification based on the memes text
(Accuracy, Precision, Recall, F1 scores)

Model	Acc	P	R	F1
Llama3.1-8B-Instruct	0.52	0.35	0.57	0.43
RoLlama3-8B-Instruct	0.62	0.34	0.20	0.25

5.4. Fake Image Detection

In this series of experiments, we are concerned with real vs fake classification of the meme image(s). Even though this is an experiment usually conducted only on the image content, we also explore the incorporation of the text content. The intuition behind this approach is that sometimes the text content may contain clues regarding the authenticity of the image. For example, considering the text *Așa ar arăta dacă...* (*This is what he would look like if...*) this may be a clear indication of a fake image. In the first series of experiments, we considered a CNN-based Residual network (ResNet50, ResNet101) for image processing. The text was processed through a DistilBERT model (the 'distilbert-base-uncased' variant, using a maximum length of 128 tokens). After extracting features from both the image and text, these features are concatenated. The combined feature vector then passes through a custom classifier. This classifier consists of a linear layer that reduces the combined features to 512 dimensions, followed by a ReLU activation and a dropout layer with a rate of 0.3. A final linear layer then produces the output for binary classification. The training process uses the following parameters (found through grid search): batch size 32, learning rate 0.0005, weight decay 1e-4, training epochs 30. For training we employ the Adam optimizer, binary cross-entropy with logits loss, and a ReduceLROnPlateau learning rate scheduler, with a patience of 3 epochs. Results are given in Table 6, in terms of macro-precision, macro-recall, macro-F1, and accuracy. The best performing model is represented by a ResNet101 image processing model, without text information. Thus, it seems that our initial assumption that the text may provide useful information for classification was wrong. Overall, it hurts the model's performance and an image-only approach is better.

Table 6: Results from CNN-based experiments for fake meme image(s) detection

Model	P	R	F1	Acc
ResNet50 + text	0.7198	0.7179	0.7187	0.7246
ResNet101 + text	0.6215	0.6293	0.6167	0.6232
ResNet50	0.7492	0.7543	0.7503	0.7536
ResNet101	0.8027	0.7727	0.7805	0.7971

A second series of experiments employed Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). We employ two architectures: CNN-based, using four convolutional layers, with max pooling, batch normalization and ReLU activation, and a pre-trained VGG16 network (Simonyan and Zisserman, 2015), with the weights of the convolutional layers frozen, allowing the rest of the layers to be trained. For training, we employed the Adam optimizer with a learning rate of 0.0001, weight decay of 1e-5, and binary cross-entropy with logits loss. A learning rate scheduler (ReduceLROnPlateau) is used to adjust the learning rate based on the validation loss, with a patience of 5 epochs and a reduction factor of 0.1. The training loop runs for a maximum of 75 epochs, with an early stopping mechanism that has a patience of 10 epochs. Gradient clipping is employed during training, with a maximum norm of 1.0, to prevent exploding gradients and stabilize the training process. This is particularly important given the depth of the network and the potential for unstable gradients in GAN-inspired architectures. For these experiments we did not include text information. Results are given in Table 7, in terms of macro-precision, macro-recall, macro-F1, and accuracy. The VGG16 results are better compared to the basic CNN approach. However, the results are lower than the previous ResNet101 approach.

Table 7: Results from GAN experiments for fake meme image(s) detection

Model	P	R	F1	Acc
GAN with CNN	0.5974	0.6000	0.5984	0.6304
GAN with VGG16	0.6693	0.6813	0.6707	0.6848

6. Conclusions and Future Work

This paper introduced the RoMemes corpus, a curated dataset of Romanian memes, with multiple levels of annotations. The dataset was publicly released. The scripts used for processing the raw corpus and for the baseline experiments are also available inside an associated GitHub repository⁹. Baseline experiments were detailed in Section 5. For all tasks, additional algorithms are needed in order to improve the overall text extraction or classification performance. Future experiments will focus on using the Romanian lexicon RoEmoLex as a replacement for the NRC and eliminating the need for translation when performing sentiment analysis. Furthermore, we'll focus on improving multimodal algorithms for the different classification levels. Of particular interest is the detection of deep fake images used in memes. With the development of generative artificial neural models, computers can generate vivid face images that can easily deceive human beings. These generated fake faces will inevitably bring serious social risks, e.g., fake news, fake evidence, and pose threats to security (Barsh et al., 2020). Our dataset will need to be expanded with additional memes containing deep fake images in order to be successfully employed for the multimodal task of deep fake detection.

References

- Avram, A.-M., Catrina, D., Cercel, D.C., Dascalu, M., Rebedea, T., Păiș, V., Tufiș, D. 2022. Distilling the Knowledge of Romanian BERTs Using Multiple Teachers. *Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association, p. 374-384.
- Barbu Mititelu, V. 2018. Modern syntactic analysis of Romanian. In Ichim, O. (coord.), Botoșineanu, L., Butnaru, D., Clim, M.R., Ichim, O., Olariu, V. (eds.), *Clasic și modern în cercetarea filologică românească actuală*, Iași, Editura Universității “Alexandru Ioan Cuza”, p. 67-78.
- Barsh, A. Y. I., et al. 2020. MPG Prediction Using Artificial Neural Network. *International Journal of Academic Information Systems Research (IJASIR)* 4(11): 7-16.
- Bauchhage, C. 2011. Insights into internet memes. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.
- Beskow, D.M., Kumar, S. and Carley, K.M., 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing and Management*, 57(2), p.102-170.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., and Kegelmeyer, W.P. 2002. Smote: synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321-357.
- Chae, Y. and Davidson, T., 2023. Large language models for text classification: From zero-shot learning to fine-tuning. *Open Science Foundation*.

⁹ <https://github.com/racai-ai/RoMemes>

- Ciobotaru, A. and Dinu, L.P. 2021. RED: A novel dataset for Romanian emotion detection from tweets. *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, p. 291-300.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, p. 4171-4186, <http://dx.doi.org/10.18653/v1/N19-1423>.
- Dhole, K., Gangal, V., Gehrmann, S., Gupta, A., Li, Z., Mahamood, S., Mahadiran, A., Mille, S., Shrivastava, A., Tan, S., Wu, T., Sohl-Dickstein, J., Choi, J.D., Hovy, E., Dušek, O., Ruder, S., Anand, S., Aneja, N., Banjade, R., Barthe, L., Behnke, H., Berlot-Attwell, I., Boyle, C., Brun, C., Cabezudo, M.A.S., Cahyawijaya, S., Chapuis, E., Che, W., Choudhary, M., Clauss, C., Colombo, P., Cornell, F., Dagan, G., Das, M., Dixit, T., Dopierre, T., Dray, P.A., Dubey, S., Ekeinhor, T., Giovanni, M., Goyal, T., Gupta, R., Hamla, L., Han, S., Harel-Canada, F., Honore, A., Jindal, I., Joniak, P.K., Kleyko, D., Kovatchev, V., Krishna, K., Kumar, A., Langer, S., Lee, S.R., Levinson, C.J., Liang, H., Liang, K., Liu, Z., Lukyanenko, A., Marivate, V., Melo, G., Meoni, S., Meyer, M., Mir, A., Moosavi, N.S., Meunnighoff, N., Mun, T.S.H., Murray, K., Namysl, M., Obedkova, M., Oli, P., Pasricha, N., Pfister, J., Plant, R., Prabhu, V., Păiș, V., Qin, L., Raji, S., Rajpoot, P.K., Raunak, V., Rinberg, R., Roberts, N., Rodriguez, J.D., Roux, C., Vasconcellos, P. H. S., Sai, A.B., Schmidt, R.M., Scialom, T., Sefara, T., Shamsi, S.N., Shen, X., Shi, Y., Shi, H., Shvets, A., Siegel, N., Sileo, D., Simon, J., Singh, C., Sitelew, R., Soni, P., Sorensen, T., Soto, W., Srivastava, A., Srivatsa, K.V.A., Sun, T., Varma, M.T., Tabassum, A., Tan, F.A., Teehan, R., Tiwari, M., Tolkiehn, M., Wang, A., Wang, Z., Wang, Z.J., Wang, G., Wei, F., Wilie, B., Winata, G.I., Wu, X., Wydmanski, W., Xie, T., Yaseen, U., Yee, M.A., Zhang, J., and Zhang, Y. 2023. NL-Augmenter: A framework for task-sensitive natural language augmentation. *Northern European Journal of Language Technology*, 9(1), p.1-41.
- Dias, M. and Carla Teixeira Lopes. 2023. Optimization of Image Processing Algorithms for Character Recognition in Cultural Typewritten Documents. *J. Comput. Cult. Herit.* 16, 4, Article 77 (December), <https://doi.org/10.1145/3606705>
- Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A. and Goyal, A., 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.
- Fellbaum, C. 2005. WordNet and wordnets. In: Brown, Keith et al. (eds.), *Encyclopaedia of Language and Linguistics*, Second Edition, Oxford: Elsevier, 665-670.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems* 27, p. 2672-2680.
- Hristova, S. 2014. Visual memes as neutralizers of political dissent. *tripleC Communication Capitalism and Critique Open Access Journal for a Global Sustainable Information Society*, 12(1), p. 265-276. <https://doi.org/10.31269/triplec.v12i1.507>
- Hwang, E.J., and V. Shwartz. 2023. MemeCap: A Dataset for Captioning and Interpreting Memes. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 1433-1445.

- Iskandarani, M. 2015. Improving the OCR of Low Contrast, Small Fonts, Dark Background Forms Using Correlated Zoom and Resolution Technique (CZRT). *Journal of Data Analysis and Information Processing*, 3, p. 34-42
- Johann, M. 2022. Political participation in transition: Internet memes as a form of political expression in social media. *Studies in Communication Sciences*, 22(1), p. 149-164. <https://doi.org/10.24434/j.scoms.2022.01.3005>
- Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. 2017. Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, p 427-431, Valencia, Spain.
- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, p. 2611-2624.
- Masala, M., Ilie-Ablachim, D.C., Dima, A., Corlatescu, D., Zavelca, M., Olaru, O., Terian-Dan, S., Terian-Dan, A., Leordeanu, M., Velicu, H. and Popescu, M., 2024. “Vorbești Românește?” A Recipe to Train Powerful Romanian LLMs with English Instructions. arXiv e-prints, pp.arXiv-2406.
- Lupea, M., and Briciu, A. 2019. Studying emotions in Romanian words using formal concept analysis. *Computer Speech and Language*, 57, p. 128-145.
- Mohammad, S., and Turney, P. 2010. Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, p. 26-34.
- Niță, S. and Păiș, V. 2024. RACAI at SemEval-2024 Task 10: Combining algorithms for code-mixed Emotion Recognition in Conversation. *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics, Mexico City, Mexico, p. 1032-1037.
- Nguyen, T.T.H., Jatowt, A., Coustaty, M. and Doucet, A., 2021. Survey of post-OCR processing approaches. *ACM Computing Surveys (CSUR)*, 54(6), p.1-37.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.J. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics*, p. 311-318.
- Parrott, W. G. (Ed.). (2001). *Emotions in social psychology: Essential readings*. Psychology Press.
- Păiș, V. and Tufiș, D. 2018. Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy Series A – Mathematics Physics Technical Sciences Information Science*. vol. 19, no. 2, p. 185-191.
- Păiș, V., Tufiș, D. and Ion, R. 2019. Integration of Romanian NLP tools into the RELATE platform. *Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing*.
- Păiș, V., Ion, R., Tufiș, D. 2020. A processing platform relating data and tools for Romanian language. *Proceedings of the 1st International Workshop on Language Technology Platforms*, p. 81-88.

- Păiș, V. 2020. Multiple annotation pipelines inside the RELATE platform. *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*. p. 65-75.
- Păiș, V., Ion, R., Avram, A.M., Mitrofan, M., Tufiș, D. 2021a. In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology*, 24(4), p. 384-401.
- Păiș, V., Barbu Mititelu, V., Irimia, E., Mitrofan, M., Gasan, C.L., Micu, R. 2022a. Romanian micro-blogging named entity recognition including health-related entities. *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop and Shared Task*, p. 190-196.
- Păiș, V., Mitrofan, M., Barbu Mititelu, V., Irimia, E., Micu, R., Gasan, C.L. 2022b. Challenges in Creating a Representative Corpus of Romanian Micro-Blogging Text. *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-10)*, p. 1-7.
- Plutchik, R. 1980. *A general psychoevolutionary theory of emotion*. *Emotion: Theory, Research, and Experience*, 1:3[33].
- Popovic, M. 2017. ChrF++: words helping character n-grams. *Proceedings of the Second Conference on Machine Translation*, Volume 2: Shared Task Papers, p. 612-618, Association for Computational Linguistics.
- Priyadarshini, I. and Cotton, C. 2022. AI Cannot Understand Memes: Experiments with OCR and Facial Emotions. *Comput. Mater. Contin.*, vol. 70, no. 1, p. 781-800.
- Schmidt, T., Hartl, P., Ramsauer, D., Fischer, T., Hilzenthaler, A. and Wolff, C. 2020. Acquisition and Analysis of a Meme Corpus to Investigate Web Culture. *15th Annual International Conference of the Alliance of Digital Humanities Organizations*, DH 2020, Conference Abstracts. Ottawa, Canada.
- Sharma, C., Paka, W., Deepesh Bhageria, S., Das, A., Poria, S., Chakraborty, T., and Gambäck, B. 2020. Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain.
- Sharma, R. S, Akhtar, M. S., and Chakraborty, T. 2024. Emotion-Aware Multimodal Fusion for Meme Emotion Detection. *IEEE Transactions on Affective Computing*, 15(3), p. 1800-1811, 10.1109/TAFFC.2024.3378698.
- Simonyan, K., and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations (ICLR 2015)*, Computational and Biological Learning Society, 2015, p. 1-14.
- Straka, M., Hajič, J., Straková, J. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.
- Tanaka, K., Yamane, H., Mori, Y., Mukuta, Y., and Harada, T. 2022. Learning to evaluate humour in memes based on the incongruity theory. *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, p. 81-93, Gyeongju, Republic of Korea. Association for Computational Linguistics.
- Tăiatu, I.M., Avram, A.M., Cercel, D.C., Pop, F. 2023. Robertweet: A bert language model for Romanian tweets. *Natural Language Processing and Information Systems*, p. 577-587, Cham. Springer Nature Switzerland

- Tufiş, D., Barbu, E., Mititelu, V., Ion, R., Bozianu, L. 2004. The Romanian Wordnet. *Romanian journal of information science and technology*, Vol 7.
- Tufiş, D., Barbu Mititelu, V., Irimia, E., Păiş, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. 2019. Little Strokes Fell Great Oaks. Creating CoRoLa, The Reference Corpus of Contemporary Romanian. *Revue roumaine de linguistique*, No./Issue 3.
- Vyalla, S. R. and Udandaraao, V. 2020. Memeify: A Large-Scale Meme Generation System. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD (CoDS COMAD 2020)*. ACM, p. 307-311.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, p. 24824-24837.
- Zacharias, E., Teuchler, M. and Bernier, B., 2020. Image processing-based scene-text detection and recognition with tesseract. arXiv preprint arXiv:2004.08079.

DO FAIRY TALES INFLUENCE THE ACQUISITION OF LANGUAGE IN PRIMARY SCHOOL? STEPS TOWARDS AN AUTOMATIC EVALUATION

MARIA UNGUREANU^{1,2}, ALEX MORUZ³,
LAURENȚIU ȘOITU^{1,4}, DAN CRISTEA^{3,5}

¹ *Doctoral School, State University of Moldova*

² *Theoretical College "Miron Costin" Pașcani*
mariamarrynna@gmail.com

³ *Faculty of Computer Science, University "Alexandru Ioan Cuza" of Iași*
alex.moruz@gmail.com

⁴ *Faculty of Psychology and Educational Sciences, University "Alexandru
Ioan Cuza" of Iași*
soitu@uaic.ro

⁵ *Institute of Computer Science, Iași branch of Romanian Academy*
dan.cristea@acadiasiasi.ro

Abstract

Language development (speech or verbal communication) is a component of education that enjoys special attention from the first years of schooling. The motivation for this concern lies in the relationship of interdependence between language and thought. The size, the variety and the use of vocabulary significantly influence the development of thinking. Compared to lexemes, phraseological units represent complex linguistic structures and their adequate integration in communication acts (oral or written) of students proves a superior intellectual capacity, because the semantic decryption of these structures involves metaphorical thinking, with all the processes defining analogy: analysis, comparison, selection, classification, interpretation. A favourable content for the training of metaphorical thinking is the fairy tale. The preponderance of phraseological units is the linguistic characteristic of this literary species. In this paper, we present a study intended to prove that the exposure of primary school students to fairy tales positively impacts the development of their linguistic performance. Our study compares the occurrence of phraseological units in students' writings before and after being exposed to fairy tales. The linguistic data of the study includes a digitised corpus of fairy tales and a collection of compositions commissioned to students enrolled in the first-cycle school. To assure diastatic generalization, the writings were collected in schools in Romania and the Republic of Moldova. A string of words was considered a phraseological unit if an occurrence of it was found in the *Dictionary of Romanian Expressions in Contexts* (DERC, author Vasile Ilincan). An initial hand-organized experiment

proved the validity of the work hypothesis. Then, for a more accurate evaluation, the phraseological units are detected by a computer program. The design of this application is described in some detail. The study concludes with an inventory of possible statistical measures and a discussion.

Keywords: fairy tales, phraseological units, development of thinking, Dictionary of Romanian Expressions in Contexts, detecting expressions, corpora

1. Introduction

The research we describe in this paper aims to identify how fairy tales, through their linguistic content in the form of phraseological units, influence the development of linguistic skills and, implicitly, the cognitive capacity of students of primary school age.

The research was carried out during the years 2021-2024 and included more phases. First, we analysed school documents (laws, regulations, curricula, methodical letters) from Romania and the Republic of Moldova, to identify the provisions regarding the use of fairy tales in instructional and educational activities. This was followed by the selection of school textbooks and their analysis from the perspective of integrating fairy tales into the learning content. Based on the findings, we determined the volumes of fairy tales that would make up the corpus for analysis. The subsequent pedagogical experiment included a training module for teachers. Over four weeks, the participants learned how to dramatize fairy tales in activities with students, and then, over two weeks, they applied this knowledge in the classroom. At the beginning and end of the exercises period, the students were evaluated. The texts created by the students and the fairy tales from the selected collections constituted the text corpus used as resources for the automatic analysis and has inspired the creation of the computer tool.

The paper has the following structure: in Section 2 we briefly present the scientific background of the research, Section 3 is dedicated to the methodology, Section 4 describes the corpora used given the automatic processing and the preliminary procedures, Section 5 shows how DERC and DEX-online have been processed to organize the counting process, Section 6 describes in some detail the program designed, Section 7 presents preliminary results and discussions, and Section 8 is dedicated to conclusions.

2. Background

The sapiential role of the fairy tale, its fixed structure (Propp, 1970) and the obligation of a conclusive ending, with emphasis on the elements to be

remembered, are meant to develop structures of thought and behaviours, superior attitudes.

As D. Tihan points out in a recent study (2017), for the research on the relationship between human language and thought there are two major antithetical paradigms: *the communicative conception* and the *linguistic relativism*. The communicative conception of language (Carruthers and Boucher, 1998) affirms the existence of a relationship of independence between language and thought. According to this postulate, language is a simple tool used by people to communicate their thoughts. In contrast, the cognitive approach states that (native) language can shape the human mind in varying proportions, resulting in the interdependence between language and thought. This theory focuses mainly on the Sapir-Whorf hypothesis (Hojer, 1954), known as relativism or linguistic determinism. A new version of linguistic relativism has crystallized, in recent years, as a result of experiments undertaken through computer simulation (Elman et al., 1996). Based on the results obtained, it was concluded that language influences the way we think about the world (Tihan, 2017). The interdependence between language and thought forces us to observe that it is a reciprocal influence. This third perspective is the one that the school values the most and constantly.

Today's society, whose communicational tendencies (we use this expression in recognition of the ample place occupied by the instruments that mediate communication between people, instead of direct communication, "face to face") are characterised by synthesis, essentializing, minimising the dimension of the utterance, imposes concerns for the development of metaphorical thinking. On this ground, we have perceived the importance of developing metaphorical thinking in children of a young age, which is substantially contributed by fairy tales, known to be extremely rich in phraseological expressions.

Complex functional phraseological units, expressions, along with locutions, occupy an important weight in the structure of any language, both at the popular and at the literary level. Such stable constructions have been established over the centuries, they have changed or gone out of use and represent testimonies of how speakers of an ethnic community think and express themselves. These expressions, merged in their contexts, most often unique, have the role to plasticize the ideas expressed and give a special flavour to communication. Their correct use, in concrete situational contexts, is a criterion for recognizing language competence and performance (Ilinca, 2015). The difference between expressions and phrases has been demonstrated by professor Theodor Hristea (1984): he calls expressions strings of words which have kept a strong expressivity (have a very strong affective load): *a spăla putina* (literary: *to wash the barrel*; approx. *to disappear, to leave quickly and unnoticed*), *a tăia frunză la câini* (lit.: *to cut leaves to dogs*; approx.: *to lose time*), *a-i lipsi (cuiva) o doagă* (lit.: *to miss*

someone a slave; approx.: *to be crazy*), *a-și lua inima în dinți* (lit.: *to take the heart in his teeth*; approx. *to dare*), etc., while phrases are groups which have lost their expressiveness or has become well-welded: *a băga de seamă* (*to notice*), *a-și aduce aminte* (*to remember*), etc.

Our research aims to highlight the extent to which children's exposure to fairy tales leaves long-lasting traces in their language. We specifically focus on identifying phraseological units, as these linguistic "tools" are defining elements of fairy tales. If our hypothesis is confirmed, it will demonstrate the importance of using this literary genre in schools for educational purposes. Currently, in Romanian schools and those in the Republic of Moldova, fairy tales occupy a limited space in school textbooks. Students' contact with fairy tales happens randomly or optionally, either in extracurricular activities or within the family. For this reason, we believe that it is necessary to revitalize discussions about the educational function of fairy tales. Foundational works could include those of psychologists such as Bruno Bettelheim (2017), Marie-Louise von Franz (2019), or more recently, articles by Elena Cocoradă (2015). Consulting the specialized literature, we found that numerous folklore studies (Hașdeu, 1972; Vrabie, 1975; Șăineanu, 1895; Roșianu, 1996; etc.), as well as some aesthetic studies (Propp, 1970; Călinescu, 1965; Danciu and Danciu, 2021) etc., make indirect references to the educational potential of fairy tales. Therefore, empirical research on this topic is necessary. Our research responds to this need, materialising in an approach based on the assumption that if students who read or listen to fairy tales acquire language in the form of phraseological units and later use them correctly in their compositions, it indicates that fairy tales can contribute to children's cognitive and emotional development by stimulating metaphorical thinking with all its defining processes.

3. Methodology

Our research followed these steps:

1. A hypothesis was formulated, stating that a child's exposure to fairy tales positively impacts the development of their cognitive abilities and writing skills.
2. Several collections of Romanian fairy tales were identified, from the entire Romanian area and Bessarabia. Only the collections of the authors Petre Ispirescu (1984), Ion Creangă (2005) and Grigore Botezatu (2005) have been included in the research, as fairy tales of these authors occur very frequently in current school textbooks in Romania and the Republic of Moldova. The collection of fairy tales is referred in the following as the Cb corpus.
3. A pedagogical experiment was organised consisting of involving students in activities to receive the fairy tale through dramatization. The experiment lasted

six weeks: four for teacher training and two for their activities with students. The students took an initial and a final evaluation, consisting of writing a fairy tale about solidarity. Thus, two distinct corpora of texts were obtained: corpus 1 (C1) includes the texts of the students from the initial test, and corpus 2 (C2) includes the texts of the students written during the final test. Their comparative analysis is thought to highlight the language acquisition as a result of the students' interaction with the fairy tale.

4. The fairy tales from the selected collections were manually analysed and findings were made regarding the frequency of phraseological units in these texts.

5. Based on the promising findings obtained by hand comparing the students' writings for inclusion of expressions known as being phraseological, a program was designed to produce a more accurate result (the actual implementation of this piece of code is still work in progress). To account for a steady classification of an expression as being phraseological, they had to exist in an inventory of such expressions.

6. All three distinct corpora, the corpus of fairy tales (Cb), and the two corpora of students' writings (C1 and C2), are now ready to be analysed with the software, and the frequency of phraseological units will be used to conclude the work (still in progress).

It is a known fact that in a phraseological expression, there is a stable part and a mobile part. Therefore, the counting process has to take this aspect in consideration, i.e. it cannot be organized as a trivial search for fixed character strings.

Examples:

sabie cu două tăișuri (lit. *double-edged sword*; approx. *a mixed blessing*) can also be used as **săbii cu două tăișuri** (lit. *double-edged swords*);

a spăla putina (lit. *to wash the barrel*; approx. *to disappear, to leave quickly and unnoticed*) can also be used with any form of the verb *a spăla*: **spălase putina**, **am spălat putina**, etc.

To account for the variability of certain items belonging to phraseological expressions, in the following we will mark them with the sign "+" (**sabie+ cu două tăișuri**, **spăla+ putina**, **lăsa+ mai moale** (lit. *to let softer*; approx. *to take it easy*)).

4. Corpus and Linguistic Resources

The corpus under analysis includes two categories of texts:

a) Literary texts: folk tales from Romania, the collection of folklorist Petre Ispirescu, which includes 36 texts, totalling 306 pages of text; folk tales from

Bessarabia, the collection of folklorist Grigore Botezatu, 45 texts in 386 pages; cult fairy tales belonging to the writer Ion Creangă, 9 texts totalling 140 pages.

b) Free compositions: 220 texts elaborated by students at the beginning and at the end of the pedagogical experiment carried out within our research.

As inventory of Romanian phraseological expressions against which to validate the occurrence of such expressions in our Cb, C1 and C2 corpora we have chosen *Dicționarul expresiilor românești în contexte* (DERC – *Dictionary of Romanian Expressions in Contexts*), authored by prof. Vasile Ilincan. DERC, published in a series of four volumes: vol. I: A-C (Ilincan, 2015), vol. II: D-N (Ilincan, 2016), vol. III: O-R (Ilincan, 2021), vol. IV: S-Z (Ilincan, 2021), is addressed to a diverse audience: pupils, students, teachers, translators, journalists and to all those interested in knowing the expressive resources of the Romanian language.

The variability of linguistic terms is given by the declension of nouns and adjectives and the conjugation of verbs. A linguistic repository including all these forms can be found in the linguistic database associated with DEX-online¹. This means that the collection of expressions given in DERC should be extended to cover the variability of flexing items.

5. Dictionary Parsing

In view of preparing the automatic processing, the original sources to be involved had to undergo transformations. As such, out of the .doc (and partially .pdf) format of DERC, as received from the author, the key expressions had to be extracted in a uniform and structured format.

The structure of the dictionary is as follows: on the top level is given the letter of the head word for the expressions; under this letter heading, the expressions as such are listed in the following form: the title expression, the definition and one or more citations from Romanian literature, each paired with the corresponding author. Here is an example (reproducing the original with respect to page arrangement):

A șterge (sau a trece) cu buretele (peste ceva) = A da ceva uitării; a ierta greșelile cuiva:

Acum trebuie să uităm. Să ștergem totul cu buretele. (ZAHARIA STANCU)

[...] și dac-ați fost în stare să treceți într-o bună dimineață cu buretele preste toată dezvoltarea istorică a țărilor, cu cât mai mult veți izbuti întru aceasta, neavând de risipit decât reparaturile făcute cu greu de conservatori. (MIHAI EMINESCU)

¹ www.dexonline.ro - permission to use it from Cătălin Frâncu and Radu Borza.

– *Să știți că eu am tras cu buretele peste tot ce s-a întâmplat [...] (ION LĂNCRĂNAN)*

Since much of the information regarding a dictionary entry is based on its formatting, the PDF files need to be converted to RTF in order to preserve it. Since the transformation to RTF introduces some supplementary new lines, the original paragraph is then recovered automatically, using heuristics, based on either capitalization or context from the previous paragraph: new paragraphs need to start with a capital letter and be preceded by either a closed bracket or a semicolon. After the text is segmented into paragraphs, we extract the entry names for all the entries, with all formatting information kept intact; this is done on the basis of the makers used for signalling the gloss start (“=” for complete entries or “v.” for synonymic entries). The gloss and examples are not necessary for our processing, and are thus discarded.

Very often, the title expression represents a compacted form of some close expressions (supposed to have a similar or very close meaning). For instance, here up, the title expression, which extends up to the “=” sign: *A șterge (sau a trece) cu buretele (peste ceva)* can be expanded in four expressions:

a șterge cu buretele peste ceva,
a trece cu buretele peste ceva,
a șterge cu buretele,
a trece cu buretele.

Moreover, the term *ceva* should be ignored in the matching process. Other expressions in titles are much more complicated. Here we reproduce one of these:

a (nu) (mai) da de (înv. în) urma (cuiva sau a ceva) sau a (nu) găsi (ori a prinde, a afla etc.) urma sau a (nu)-i da de urmă, a da de o urmă (a cuiva sau a ceva). v. urmă.

Only the items in bold contribute to the expansion, the others (like *înv.*, *sau*, *ori*, *v.*, *etc.*) are linking words and, generally, indicate variations of the principal expression. The first operation in the expansion is to eliminate the infinitive particle “a” in front of verbs (as the infinitive is not the only accepted form in texts). Their inclusion could foul the counter in all cases in which the verb occurrence is in another tense than infinitive (the miss to find one item of an expression in the text will invalidate the possible occurrence). Then the title expression is split according to the surface (i.e., not included in parentheses) items *sau*, *ori* or comma (.). As such, the entry title above has to be detailed in the following four derived expressions (after elision of the particle “a”):

(nu) (mai) da de (înv. în) urma (cuiva sau a ceva)

(nu) găsi (ori a prinde, a afla etc.) urma

(nu)-i da de urmă

da de o urmă (a cuiva sau a ceva)

We try below an expansion of the first of these group expressions, by following a depth-first search ordering:

nu (mai) da de (înv. în) urma (cuiva sau a ceva) =>

nu mai da de (înv. în) urma (cuiva sau a ceva) =>

nu mai da de urma (cuiva sau a ceva) => nu mai da de urma ⇐

nu mai da în urma (cuiva sau a ceva) => nu mai da în urma ⇐

nu da de (înv. în) urma (cuiva sau a ceva) =>

nu da de urma (cuiva sau a ceva) => nu da de urma ⇐

nu da în urma (cuiva sau a ceva) => nu da în urma ⇐

(mai) da de (înv. în) urma (cuiva sau a ceva) =>

mai da de (înv. în) urma (cuiva sau a ceva) =>

mai da de urma (cuiva sau a ceva) => mai da de urma ⇐

mai da în urma (cuiva sau a ceva) => mai da în urma ⇐

da de (înv. în) urma (cuiva sau a ceva) =>

da de urma (cuiva sau a ceva) => da de urma ⇐

da în urma (cuiva sau a ceva) => da în urma ⇐

So, the first of the four groups has 8 extensions (marked with the ⇐ sign), which is $2^{**}3$, because there are 3 binary extension seeds in this group. In all, this entry extends to 21 forms, considering only the infinitive form of the main verb, ***a da*** (to give). This verb has approximately 60 forms in conjugation, which make a total of approximately 1200 forms for this entry, truly one of the most prolific in DERC. But in other expressions, there is more than one element possible to have lexical variations, so we are not far from the truth that if we want to appreciate the total number of expressions possible to occur in language, we have to multiply by 1000 the dimension of the dictionary in the number of expressions entries. Since an estimation of the number of entries in DERC is between 70,000 and 80,000, the total number of forms of expressions could reach 80 million.

The DERC dictionary entries, processed as described above, were saved under the name Expanded-DERC. We are aware that not all words in an expression that have morphological variations (therefore, belonging to the categories: verb, noun, adjective and pronoun) are allowed to vary in real usage, but this proliferation of

forms in the Expanded-DERC repository makes no harm (by producing false positives) because the language use will simply not include them.

So, to resume, after the dictionary is transposed to plain text, definitions and citations have to be removed and the remaining entries' titles have to be expanded by using the auxiliary items that indicate variants. Below is a sketch of the Expand-DERC algorithm, whose input is a primary title entry of DERC and whose output is the expansion of that entry, as discussed above. Iterated on all entries in DERC, the algorithm is supposed to produce the Expanded-DERC repository.

```

SOE = set_of(<title_string, 0>);
; SOE = set of embryos; we call embryo a pair consisting of a string of
items and an index into that string; initially, the string consists of the dictionary
entry, and the index points to the first word of the string
; title_string = the string in the title of the dictionary entry
; 0 = index into this string from where to continue parsing; if the index
has exceeded the last item of the string, it takes the value 1000 (we consider there
are no such long strings)
while (there are still indices other than 1000 in
SOE)
  <string, index> := first_non1000_of(SOE);
  ; choose in SOE an embryo that has not been fully parsed
  delete <string, index> from SOE;
  ; eliminate the selected embryo from SOE;
  if (index >= length_of(string) then index := 1000;
  else
  while (string[index] =/ parenthesis)
  index++;
  ; on the index position in string there is a parenthesis
  elem := inside_par(string);
  ; let elem be the string in the parenthesis
  if length(elem)=1 then
    ; include in SOE: <string without elem, index>;
    ; include in SOE: <string in which elem is not
boarded by parenthesis, index++>;
    else if (first(elem) == "înv.") then
    ; include in SOE: <string without elem, index>;
    ; include in SOE: <string in which elem is replaced
by second(elem), index++>;
    else if (first(elem) == "ori") then
    ; case of a parenthesis starting with "ori"
    include in SOE: <string without elem,
index>;

```

```

                                else if (elem does not contain ",")
then
    ; case of a parenthesis without enumeration
                                include in SOE: <string in which the item on the
                                position index-1 is replaced by rest(elem),
                                index>;
                                else
    ; case of a parenthesis with enumeration
                                include in SOE: <string in which the item on the position index-1
is replaced by till_comma(elem) + "(" + after_comma(elem) +
    ") ", index>;
                                print(strings_of(SOE));
                                ; when there are no more unparsed embryos in SOE, output all elements
in SOE

---List of mentioned functions ---
first_non1000_of(SOE) <= returns the first embryo in SOE whose
index is not 1000
inside_par(string) <= string is a parenthesis; returns the
string inside it
first(elem) <= first item of the string elem
second(elem) <= second item of the string elem
rest(elem) <= the rest of elem after the first item
till_comma(elem) <= returns the substring of elem till the first
comma
after_comma(elem) <= returns the rest of elem after the first
comma
strings_of(SOE) <= returns all (simple) strings of SOE

```

To resume, initialised with the entry's title expression, the algorithm maintains a list of partially parsed strings and at each multiplication item (*sau, ori, comma, etc.*) the string parsed up to the index point is first removed and then forked into two other derived strings. An index indicates the position in each string from where the parsing should go on. The activity fades down when there are no more partially parsed strings in the list.

6. Processing

In order to perform the planned statistics over the corpus, the students' raw texts had to be cleaned and formatted as plain text files. Then, a pattern-matching process has to be run over the corpus. The pattern-matcher compares the students'

texts against the Expanded-DERC, counting each occurrence of an expression. The matcher allows up to three supplementary words in-between two consecutive items of an expression. This is necessary in order to allow variations of the expressions, as here: *a dat [în sfârșit] de urma* (lit. *to give [finally] of the trace; to [finally] trace*), *a dat [într-un târziu] de urma* (lit. *give [eventually] of the trace; to trace [later on]*), where the strings *în sfârșit* (*finally*), *într-un târziu* (*later on, eventually*) are not part of the expressions. Note that the string *dăduse de [o] urmă* (*had found [a] clue*) is not an expression. The confusion is generated by considering the noun *urma* as variable. To eliminate cases like this one, the variable items in each entry of Expanded-DERC, as explained in Section 3, should carry a distinguishing mark, signalling that a certain item could occur in any of its morphological forms, as here: *nu mai da+ de urma*, *nu mai da+ în urma*, *nu da+ de urma*, *nu da+ în urma*, *mai da+ de urma*, *mai da+ în urma*, *da+ de urma*, *da+ în urma* (while *urma* should not carry this mark). This type of marking necessitates a very attentive manual annotation of the whole dictionary, which is work in progress right now. Without these markings the pattern-matcher would validate in texts all morphological variations of verbs, nouns, adjectives and pronouns in the title expression, without exception, which could induce a proliferation of possible many false expressions.

Essentially, the pattern-matcher compares the input text against the transformed strings in the Expanded-DERC and increments a counter each time a match is recognised.

7. Data, Results, Statistics and Discussions

The hand-performed preliminary analysis, as said already, showed promising results. In all, we used 20 texts in Cb, from which we selected 20 of the most common phraseological units. Out of the expressions found in this corpus, 4 were also recognized in the C1 corpus and 10 in the C2 corpus, showing a net increase of 6 expressions (see Table 1, in which the intersections Cb-C1 and Cb-C2, over the entire corpus, are highlighted).

Table 1: High-frequency phraseological units in the analysed corpora

Nr. crt.	High-frequency phraseological units in Cb	High-frequency phraseological units in C1	High-frequency phraseological units in C2
1	a fost odată ca niciodată	a fost odată ca niciodată	a fost odată ca niciodată
2	a-și lua ziua bună	a cere voie	a băga de seamă

Nr. crt.	High-frequency phraseological units in Cb	High-frequency phraseological units in C1	High-frequency phraseological units in C2
3	a-și aduce aminte	a -și lua ziua bună	a-și aduce aminte
4	a da în genunchi	a avea noroc	a-și încerca norocul
5	a-și lua lumea-n cap	a băga de seamă	a o lua la sănătoasa
6	a băga de seamă	a da binețe	a-și bate joc
7	a-i crește inima de bucurie	a-și lua lumea-n cap	a da curaj
8	a se lumina de ziuă	a izbucni în plâns	a prinde curaj
9	a-și bate joc	a sta pe gânduri	a-și lua inima-n dinți
10	a pune masa	a face un plan	a-și găsi un rost
11	a-și lua inima în dinți	a avea nevoie de ceva/cineva	a o lua la sănătoasa
12	a se prinde în horă	a sta pe gânduri	a da binețe
13	a merge în peștit	nici una nici două	a merge în peștit
14	a-și pune în gând	a pune la inimă	a-și încerca norocul
15	a ține ochii la ceva/pe ceva/cineva	a-și găsi un rost	de-ale gurii
16	a pune/a sta pe gânduri	a întinde masă	a-și lua ziua bună
17	a da curaj	a prinde cu minciuna	a cere mâna (cuiva)
18	a se pierde cu firea	a-și face curaj	a sta pe gânduri
19	a o lua la sănătoasa	a râde de ceva/cineva	a muri de foame
20	a o rupe la fugă	a da dreptate	a purta gândul

To limit the researcher's subjectivity (which was the first author), only the expressions found in DERC (Ilinca, 2015; 2017; 2021a; 2021b) were recorded. In the table, columns 1, 2 and 3 show the first 20 most frequent expressions in Cb, C1 and C2, respectively. A certain intersection between Cb and C2 can be remarked, which certifies the influence of fairy tales on the formation of children's high vocabulary. Comparing the figures that show the total number of recognized expressions in C1 and C2, a similar conclusion can be inferred.

At this moment we do not have the final results, after the implementation of the software to process the linguistic resources and the corpora (as said, this is still work in progress), but we do not expect a sharp discrepancy in the results.

Other statistics that could be easily handled having the described software would be:

- a comparison between the growth of expressions acquired by students per region, respectively in the Republic of Moldova and Romania. If $T1M$, $T2M$ = the total number of expressions counted in C1 and C2 in Moldova, and $T1R$, $T2R$ = similarly for Romania, compare $T2M - T1M$ versus $T2R - T1R$;
- a comparison between the quantity of fairy tales given to students and their progress in acquiring language; in other words, is the quantity correlated with the acquisition of language? Suppose there are two groups of students, $G1$ and $G2$, $Q1$ is the quantity of texts exposed to $G1$ and $Q2$ - same for $G2$, $T21$ is the number of expressions in C2 of $G1$ and $T22$ is the number of expressions in C2 for $G2$, if $Q1 > Q2$ is it that $T21 > T22$?
- is there a correlation between the frequency of an expression in the exposed tales and its use by children? Therefore, is it true that a frequently cited expression asserts a better stabilization in the language of children? Then, can differences be distinguished among expressions themselves, some being more easily integrated/acquired in language use than others?

8. Conclusions

The students adopted the structure of the fairy tale in their own works. They noticed that the structural elements in the expression reflect a goal-oriented way of thinking, and they took over in the texts the need for a "happy" ending for both the heroes of the story and the community (all the lands of the neighbouring districts participate in the wedding). The vocal expression is much more elaborate, because it has a "special" audience, eager to learn, to enjoy toiling together (no one enters the story without having a certain role).

In the future, we want to refine all these studies and extend them to more groups of students. A more refined analysis can now be drawn, which would consider the peculiarities of expressions themselves, not only their net occurrence. If particular groups of children of different ages are also selected by voluntary agreement and specially chosen fairy tales are indicated to them, one can also develop differentiated conclusions concerning the openness of students to enriching their language depending on the age and content of the text they have been exposed to.

Moreover, with these pieces of software built, the study can now be extended to other types of texts as well, far away from strictly fairy tales. The idea is that we can now prove with concrete figures that readings improve the acquisition and

the use of language from a fairly young age till maturity, and what are the dynamics of this very complicated process.

Finally, our research enriches the public domain linguistic resources, because we have already the agreement of the proprietary of DERC to offer this resource for research purposes. To the original electronic version of DERC, we can now add the Expanded-DERC resource, a combination between a transformed version of DERC and the lexical database of DEX-online. In time, this resource could be enhanced by eliminating variations of items that should be considered frozen in expressions (see a remark on this issue in Section 6).

Acknowledgements

We are much indebted to professor Vasile Ilinca, who offered us the electronic version of his dictionary to accomplish this research. He also agreed to place the Expanded-DERC electronic resource on the public domain for research purposes.

References

- Bettelheim, B. 2017. *Psihanaliza basmelor*, (EN: *Psychoanalysis of Fairy Tales*), Editura Univers, București.
- Botezatu, G. (2005), *Povești populare din Basarabia* (EN: *Folk tales from Bessarabia*), Editura Prut Internațional, Chișinău.
- Carruthers, P. and Boucher, J. 1998. *Language and Thought: Interdisciplinary Themes*. Cambridge University Press.
- Călinescu, G. 1965. *Estetica basmului* (EN: *The aesthetics of the Fairy Tale*), Editura pentru literatură, București.
- Cocoradă, E., (2015) *Basmul – tărâm al paradoxului* (EN: *The Fairy Tale – A Realm of Paradox*), in Liliana Ezechil-60, coord. Langa, C, Soare, E., Editura Paralela 45, Pitești.
- Creangă, I. (2005), *Amintiri din copilărie. Povești. Povestiri* (EN: *Childhood memories. Fairy Tales. Stories*). Editura Litera Internațional, București.
- Danciu, L., Danciu, A.P. 2021. *Zmei, draci și strigoi. Eseuri de hermeneutică a basmului popular românesc* (EN: *Dragons, Devils and Ghosts. Essays on the Hermeneutics of the Romanian Folk Tale*), Editura Universitară, București.
- Elman, J., Bates, E., Johnson, M., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. 1996. *Rethinking innateness: A connectionist perspective on development*. MIT Press/Bradford Books, Cambridge.
- Franz, M. L. 2019. *Interpretarea basmelor* (EN: *The Interpretation of Fairy Tales*). Editura Trei, București.
- Hașdeu, B.P. 1972. *Etymologicum Magnum Romaniae*, Editura Minerva, București.
- Hristea, Theodor (1984). *Sinteze de limbă română* (EN: *Synthesis of Romanian Language*). 3rd edition, Editura Albatros, Bucharest.
- Holger, H. 1954. *Language in culture: Conference on the interrelations of language and other aspects of culture*. University of Chicago Press.
- Ilinca, Vasile (2015), *Dicționar de expresii românești în contexte* (DERC) (EN: *Dictionary of Romanian Expressions in Contexts*), A-C, vol. I, Editura Presa Universitară Clujeană, Cluj-Napoca.

- Ilincan, Vasile (2016), *Dicționar de expresii românești în contexte* (DERC), D-.N vol. II, Editura Presa Universitară Clujeană, Cluj-Napoca.
- Ilincan, Vasile (2021), *Dicționar de expresii românești în contexte* (DERC), O-R, vol. III, Editura Presa Universitară Clujeană, Cluj-Napoca.
- Ilincan, Vasile (2021), *Dicționar de expresii românești în contexte* (DERC), S-Z, vol. IV, Editura Presa Universitară Clujeană, Cluj-Napoca.
- Ispirescu, P. (1984), *Legende sau basmele românilor* (EN: *Legends and Fairy Tales of Romanians*), Editura Facla, Timișoara.
- Propp, V.I. 1970. *Morfologia basmului*. Editura Univers, București.
- Roșianu, N. 1997. *Poetică folclorică* (EN: *Folk poetics*), Editura Universității din București.
- Șăineanu, L. 1895. *Basmele române* (EN: *Romanian Fairy Tales*), Tipografia Carol Göbl, București.
- Tihan, D. 2017. *Despre relația dintre limbaj și gândire – scurtă introducere în sociolingvistică* (EN: *On the relationship between language and thought – a brief introduction to sociolinguistics*). In *Social Innovation Magazine*, year IX, no. 1.
- Vrabie, Gh. 1975. *Structura poetică a basmului* (EN: *The Poetic Structure of Fairy Tale*), Editura Academiei Republicii Socialiste România, București.
- .

EXPLORING LEGAL LINGUISTIC PATTERNS: A CORPUS-BASED ANALYSIS OF ROMANIAN LAW TEXTS

ALEXANDRU DINU

*Faculty of Electronics, Telecommunications and Information Technology,
National University of Science and Technology Politehnica Bucharest, 061071
Bucharest, Romania
alexandru.dinu2408@upb.ro*

Abstract

Understanding and managing linguistic data play crucial roles across various domains. Among these, the legal field stands out due to the diversity and complexity of its language. In this context, the project presented in this paper aimed to create a representative linguistic corpus for written Romanian, with a specific emphasis on the language used in legal document drafting. To achieve this goal, we used the concept of a representative corpus and analyse the key dimensions involved in its construction. These dimensions include probabilities associated with n-grams or individual words, conditional probabilities, and informational entropy. Despite previous attempts to research and analyse linguistic corpora in the legal domain, this topic remains insufficiently addressed. Therefore, our project seeks to fill this gap by developing and analysing a representative legal linguistic corpus for the Romanian language. Our primary objective is to identify potential similarities and differences between this legal corpus and a general literary corpus. To achieve this objective, our project involves comparison of linguistic entities such as words, phrases, or n-grams between the two corpora. We expect similarities between the corpora, with any differences in linguistic entities being exemplified statistically and qualitatively. In summary, this project contributes to a deeper understanding of legal language by constructing and analysing a representative legal corpus in Romanian, shedding light on linguistic nuances and patterns within legal texts.

Keywords: legal corpus, statistical methods, entropy, representative corpus, natural language processing

1. Introduction

A linguistic corpus consists of a large collection of written or oral texts, systematically organized to facilitate linguistic analysis. It plays a key role in fields such as language studies, machine translation, and machine learning.

Linguistic corpora contain authentic texts, providing a solid foundation for statistical analyses and comparisons. They enable the study of word frequencies, grammatical patterns, regional variations, and dialects, being essential for developing NLP technologies, such as machine translation, speech recognition, and speech synthesis.

In the legal domain, corpora are crucial for analysing and applying laws, offering extensive access to data derived from official documents. They enable comparative studies of legislative texts, facilitating the understanding of the evolution of legal language and trends. A legal corpus can help track changes in legal language over time and provide deeper insights into specialized terms found in judicial texts (Sferle, 2005).

Legal corpora are essential for the development of technologies in the legal field, such as search systems and machine translation, facilitating access to updated information and streamlining research (Biber et al., 1998; McEnery and Wilson, 2001). They enable the analysis of legal discourse (Tognini-Bonelli, 2001; Baker, 2006), the evaluation of language influence on judicial decisions, and the coherence of legal interpretation over time. Additionally, legal corpora support law students' training, helping them develop analytical skills essential for effective legal practice.

Prior studies on legal linguistic patterns in other languages have shown similar findings regarding the structured and repetitive nature of legal language, in contrast to the more flexible and creative structure typically found in literary texts. For example, works like Steinberger et al. (2006) and (eur, 2024) analyse the standardized structure of legal corpora, which often emphasizes consistency and clarity over stylistic diversity, providing insights into the unique linguistic properties of legal documents compared to literary texts.

The first Romanian linguistic corpora emerged in the 1970s and 1980s, with *Cuvinte Românești* being one of the most prominent, primarily utilized for dictionary creation and linguistic research (Sala, 2006). With technological advancements, larger and more annotated corpora emerged, such as "RoWaC" (Macoveiciuc and Kilgarrieff, 2010), reflecting the language used online, and CoRoLa, a reference corpus of contemporary Romanian, covering various domains and time periods. The CoRoLa project, initiated in 2014, is publicly accessible and adheres to international standards (cor, 2014). These corpora are used for developing NLP technologies, machine translation, and linguistic research.

Legal corpora are essential tools in legal analysis and research, providing access to laws, court decisions, and other official documents. They include linguistic and legal annotations and are used for legislative studies, analysing legal changes, and developing automated legal assistance systems. Notable projects like JRC-Acquis

(Steinberger et al., 2006), Eur-Lex (eur, 2024), and LII (cor, 2024) provide access to legislative data from Europe and the U.S., contributing to the development of legal terminology and translation. Creating a corpus involves collecting and processing data from official sources, with applications including case analysis and legal document generation. The current project aims to create a representative legal corpus for Romanian, continuing research in the field.

This paper investigates the construction and analysis of a representative legal corpus for the Romanian language, emphasizing its comparison with a general literary corpus. The paper is organized as follows:

Section 1 - Introduction: This section introduces the importance of linguistic corpora in various domains, emphasizing the complexity of legal language and the need for a representative corpus to support linguistic research and NLP development in the legal field.

Section 2 - Legal Corpus Creation: This section details the process of constructing the legal corpus, including the selection of legal texts and the steps taken to preprocess the data for linguistic analysis.

Section 3 - Experimental Results: This section presents the results of the analysis, focusing on word frequencies, letter occurrences, and bigram patterns within the legal corpus, as well as comparisons with the literary corpus.

Section 4 - Conclusions: The final section summarizes the main findings of the study, discussing the implications of the results for future research and NLP applications in the legal domain.

Each section is designed to progressively build an understanding of the linguistic properties of legal texts and how they differ from or resemble literary language, providing insights into both domains.

2. Legal Corpus Creation

In this project, we aimed to create a representative corpus that captures the language used in Romanian legal texts. The corpus selection criteria focused on ensuring diversity across major legal subfields – civil, criminal, labour, and fiscal law. By including these domains, the corpus encompasses a variety of linguistic structures and terminology specific to different types of legal documents. 21 current Romanian legal texts were used, which were then read into a script developed in MATLAB and concatenated into a single text to be used in the following stages. The texts were originally in Word format but were converted to .txt format to facilitate working with these documents in the MATLAB environment. The end result was a legal text of more than 7 million characters and 1 million words, out of which approximately 90,000 are unique terms, capturing

the specialized vocabulary of the legal field. In contrast, the literary corpus included approximately 6 million words and 200,000 distinct words.

The documents used for forming the legal corpus are as follows:

1. "Civil Procedure Code from 2010 - Republished" (Parliament, 2010a)
2. "Fiscal Procedure Code from 2015 (Law no. 207/2015)" (Parliament, 2015a)
3. "Criminal Procedure Code from 2010 (Law no. 135/2010)" (Parliament, 2010b)
4. "Fiscal Code from 2015 (Law no. 227/2015)" (Parliament, 2015b)
5. "Labor Code from 2003 (Law no. 53 of January 24, 2003) - Republished" (Parliament, 2003)
6. "Criminal Code from 2009 (Law no. 286/2009)" (Parliament, 2009)
7. "Customs Code from 2006 of Romania" (Parliament, 2006a)
8. "Constitution of 1991 - Republished" (Parliament, 1991)
9. "Decision 394/2016 approving the Methodological Norms for the application of the provisions regarding the award of the sectoral contract/framework agreement from Law no. 99/2016 on sectoral procurement" (Government, 2016a)
10. "Decision 395/2016 approving the Methodological Norms for the application of the provisions regarding the award of the public procurement contract/framework agreement from Law no. 98/2016 on public procurement" (Government, 2016b)
11. "Law 8/1996 on copyright and related rights - Republished" (Parliament, 1996)
12. "Law 62/2011 on social dialogue - Republished" (Parliament, 2011)
13. "Law 95/2006 on healthcare reform - Republished" (Parliament, 2006c)
14. "Law 99/2016 on sectoral procurement" (Parliament, 2016c)
15. "Law 100/2016 on concessions for works and services" (Parliament, 2016a)
16. "Law 101/2016 on remedies and appeals in public procurement, sectoral contracts, and concessions, as well as for the organization and functioning of the National Council for the Settlement of Appeals" (Parliament, 2016b)
17. "Law 273/2006 on local public finances" (Parliament, 2006b)
18. "Law 500/2002 on public finances" (Parliament, 2002)
19. "Law 544/2001 on free access to information of public interest" (Parliament, 2001)
20. "Law 554/2004 on administrative litigation" (Parliament, 2004)
21. "Emergency Ordinance 57/2019 on the Administrative Code" (Government, 2019).

We selected documents spanning diverse themes to create a corpus that encompasses various legal subfields, including civil procedure, criminal procedure, labour law, and fiscal law, thereby ensuring balanced representation within the legislative landscape.

Data processing in the project was carried out in several stages with the goal of obtaining a compact text composed only of words separated by spaces, making it easier to use in subsequent procedures. Initially, non-alphabetic characters were defined and replaced with spaces. Then, numbers were also removed and replaced with spaces. However, hyphens were left alongside letters, as some occur within words, and removing them at this stage would lead to the loss of the meaning and form of certain words. Therefore, it was necessary to distinguish between hyphens inside words and those outside of them to keep the words intact. Thus, hyphens not surrounded by letters were defined and replaced, like the non-alphabetic characters and numbers, with spaces. To streamline the analysis, punctuation marks were removed from the corpus. This decision was made to ensure that the analysis focused solely on the linguistic patterns of words and word sequences without interference from punctuation symbols, which are often highly variable. By excluding punctuation, we were able to create a cleaner dataset that highlights word frequencies and bigram patterns relevant to the legal language.

After applying these operations, only words and spaces remained in the unified text. Due to multiple character replacements with spaces, some words could be separated by multiple spaces, resulting in an uneven distribution in the corpus. The solution was to identify sequences of multiple spaces and reduce them to a single space.

The final text was segmented into words, which were stored in a vector. For each word, the number of occurrences and its frequency were calculated and stored for future statistical analyses.

Next, the text was segmented into letters. The first step was defining the alphabet, followed by calculating the number of occurrences and frequency for each letter. These values were also stored in vectors for statistical processing.

Another step was grouping the words into digrams, meaning pairs of consecutive words. As with letters and words, the number of occurrences and frequency for each digram were calculated and stored for future mathematical analyses.

3. Experimental Results

3.1. The Legal Corpus

To better visualize the data, a file was created for each unit type (words, letters, and digrams), each containing three columns. The first column specifies the type of unit (letter, word, or digram), the second column contains the number of occurrences in the legal corpus, and the third shows their frequencies. These results can be viewed in spreadsheet software. Table 1 contains the top 20 most frequent results for unique words in the legal corpus.

Table 1: Word occurrences and frequencies in the legal corpus (in percentages)

Word	Occurrences	Frequency (%)	Word	Occurrences	Frequency (%)
de	78014	6.53	pentru	11558	0.97
în	46394	3.89	din	12738	1.07
și	32766	2.74	alin	9793	0.82
a	29869	2.50	prin	7690	0.64
la	27835	2.33	nu	7188	0.60
care	19028	1.59	pe	7059	0.59
sau	18164	1.52	al	6549	0.55
art	16472	1.38	este	6549	0.55
se	15728	1.32	să	6535	0.55
cu	14049	1.18	o	5822	0.49

To enhance data visualization and manage the extensive table size, we focused on the top 20 words, sorted in descending order based on their frequency, which also correlates with the number of occurrences in the corpus. It can be observed that the highest frequency is associated with the word “de,” followed mostly by other connecting words, which are very common in both written and spoken Romanian.

Following this, we conducted a statistical analysis of letter frequencies within the legal corpus. As seen in Table 2, the data presents the occurrences and frequency of each letter in the legal corpus. The “Occurrences” column shows how many times each letter appears in the text, while “Frequency” represents the proportion of the total corpus that each letter accounts for. For instance, the letter “e” appears 873,798 times, making up 12.46% of the entire corpus, the highest of all letters. In contrast, rare letters like “q”, “w”, and “y” have very few or no occurrences, with a frequency of 0%.

Table 2: Occurrences and frequencies of letters in the legal corpus (in percentages)

Letter	Occurrences	Frequency (%)	Letter	Occurrences	Frequency (%)
E	873798	12.46	Ț	100848	1.44
I	804499	11.41	V	81070	1.16
A	693049	9.88	F	77322	1.10
R	585672	8.35	Z	71004	1.01
T	479842	6.84	Î	69285	0.99
n	437206	6.23	B	67488	0.96
u	400312	5.71	G	55689	0.79
L	388313	5.54	Ș	48513	0.69
C	364770	5.20	X	19115	0.27
o	310293	4.42	J	16719	0.24

Letter	Occurrences	Frequency (%)	Letter	Occurrences	Frequency (%)
S	249842	3.56	H	14607	0.21
d	246315	3.51	K	575	0.01
p	231058	3.29	W	118	0.01
Ă	177811	2.53	Y	84	0.01
m	138792	1.98	Q	57	0.01

As seen in Table 3, the data illustrates the occurrences and frequencies of the most common word digrams (pairs of consecutive words) found in the legal corpus. The "Occurrences" column shows how many times each digram appears, while "Frequency" represents the proportion of the entire corpus that the digram accounts for. For example, the most frequent digram, "in care," appears 4,776 times, making up 0.4% of the entire corpus. Other frequent digrams, such as "art alin" and "de la," also have high occurrences, indicating their common usage in legal texts. These digrams often consist of functional words or terms related to legal structuring, reflecting the standardized nature of legal language. The lower-frequency digrams still contribute to the overall structure and flow of the legal text but occur less frequently within the corpus.

Table 3: Occurrences and frequencies of digrams in the legal corpus (in percentages)

Digram	Occurrences	Frequency (%)	Digram	Occurrences	Frequency (%)
in care	4776	0.40	la data	2226	0.19
art alin	4649	0.39	a fost	2060	0.17
de la	4562	0.38	în condițiile	2009	0.17
în cazul	4040	0.34	cazul în	1984	0.17
prevăzute la	3874	0.32	alin lit	1938	0.16
la art	3294	0.28	după caz	1900	0.16
de către	3252	0.27	care se	1898	0.16
la alin	2939	0.25	și a	1632	0.14
precum și	2830	0.24	până la	1474	0.12
de a	2335	0.20	se aplică	1435	0.12

3.2. Comparing Legal vs Literary Corpora

The experiments are conducted on a literary corpus (Dinu, 2020a; Dinu, 2020b; Dinu and Vlad, 2022; Hanu et al., 2016; Hanu et al., 2018; Vlad et al., Iunie 2020; Vlad et al., Noiembrie 2020; Dinu and Vlad, 2021; Dinu et al., 2020b; Dinu et al., 2019; Hanu et al., 2019; Dinu et al., 2020a) comprising over 6.3 million words, drawn from 49 books by 9 authors. The collection includes works by Isaac Asimov (9 books), Constantin Chiriță (5 books), Alexandre Dumas (12 books), Colin

Falconer (1 book), Frank Herbert (8 books), Larry Niven (1 book), Orson Scott Card (3 books), Michel Zevaco (7 books), and J.R.R. Tolkien (3 books). This corpus has been previously used in various analyses of the authors. The corpus operates with an alphabet of 31 symbols, consisting of the basic set of letters: A, Ă, Â, B, C, D, E, F, G, H, I, Î, J, K, L, M, N, O, P, Q, R, S, Ș, T, Ț, U, V, W, X, Y, Z, and a blank space. A word is defined as any sequence of characters between two spaces or blanks.

Table 4 presents a comparative analysis of common words between the legal and literary corpora. It includes both the *number of occurrences* and the *frequency* (in percentages) for each word in both corpora. The words are listed in descending order based on their frequency in the legal corpus. Some key observations are as follows:

Words like “*de*”, “*în*” and “*și*” appear frequently in both corpora, reflecting their common usage across different types of texts.

Certain words, such as “*art*” and “*alin*”, have significantly higher occurrences in the legal corpus due to their specific legal connotations, while they are almost absent in the literary corpus.

On the other hand, words like “*o*” and “*nu*” show a much higher frequency in the literary corpus, which might indicate a stylistic or contextual difference between legal and literary texts.

Table 4: Results of common word analysis between the legal corpus and the literary corpus

Word	Occurrences_legal	Occurrences_literary	Frequency_legal (%)	Frequency_literary (%)
de	78014	221173	6.53	3.57
în	46394	132429	3.89	2.13
și	32766	165803	2.74	2.67
a	29869	50914	2.50	0.82
la	27835	77911	2.33	1.26
care	19028	61132	1.59	0.99
sau	18164	13823	1.52	0.22
art	16472	9	1.38	0.01
se	15728	101218	1.32	1.63
Cu	14049	84550	1.18	1.36
din	12738	48233	1.07	0.78
Pentru	11558	31235	0.97	0.50
alin	9793	4	0.82	0.01
prin	7690	9759	0.64	0.16
Nu	7188	93624	0.60	1.51
pe	7059	91906	0.59	1.48
al	6549	13696	0.55	0.22
este	6549	13997	0.55	0.23
să	6535	125886	0.55	2.03
O	5822	88244	0.49	1.42

Table 5 compares the frequency of each letter between the legal and literary corpora. It provides both the *number of occurrences* and the *frequency percentage* for each letter in both corpora. Some notable findings include:

Common vowels like “e”, “i”, and “a” dominate both corpora in terms of frequency. Letters such as “ä” and “ş” show a significantly higher frequency in the literary corpus compared to the legal corpus, which may reflect linguistic or stylistic differences. On the other hand, letters like “x” and “q” have minimal usage in both corpora, with very low or zero frequency in both.

This comparative analysis (Table 5) provides insights into the distribution of letters across different types of texts, highlighting patterns that may be tied to the specific nature of legal versus literary language.

Table 5: Results of letter frequency comparison between the legal and literary corpora

Letter	Occurrences_ legal	Occurrences_ literary	Frequency_ legal (%)	Frequency_ literary (%)
E	873798	3357976	12.46	11.58
I	800499	2801512	11.41	9.66
A	693049	2846564	9.88	9.82
R	585672	1996360	8.35	6.89
T	479842	1727324	6.84	5.95
n	437206	1860943	6.23	6.49
u	400312	1838985	5.71	6.42
L	388313	1279970	5.54	4.42
C	364770	1512718	5.20	5.22
o	310293	1156733	4.42	3.99
S	249842	1348243	3.56	4.65
d	246315	1018456	3.51	3.52
p	231058	929326	3.29	3.21
Ä	177811	1240124	2.53	4.29
m	138792	903253	1.98	3.12
Ť	100848	288714	1.44	0.99
v	81070	384880	1.16	1.32
F	77322	338111	1.10	1.17
Z	71004	194291	1.01	0.67
Î	69285	405918	0.99	1.40
b	67488	295783	0.96	1.02
g	55689	285039	0.79	0.98
Ş	48513	416264	0.69	1.43
Ä	19115	263756	0.27	0.91
x	16719	2835	0.24	0.01
J	14607	68022	0.21	0.24

Letter	Occurrences_ legal	Occurrences_ literary	Frequency_ legal (%)	Frequency_ literary (%)
h	14368	136124	0.21	0.47
k	575	11745	0.01	0.04
q	57	3753	0.01	0.01
w	118	8994	0.01	0.03
y	84	17908	0.01	0.06

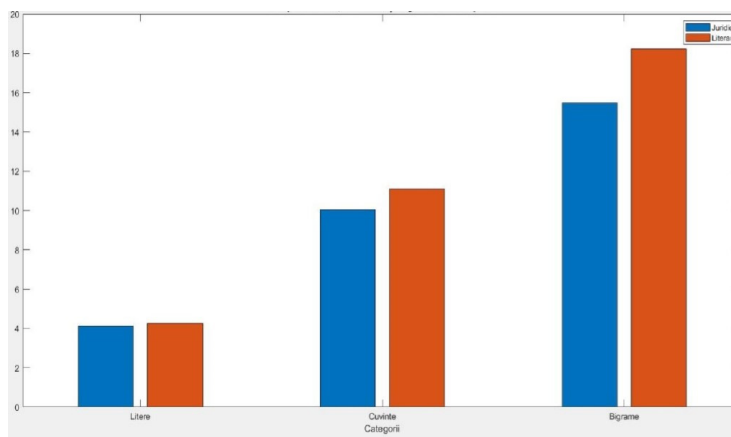


Figure 1: Entropy comparison results for letters (left), words (middle) and word digrams (right)

Figure 1 presents a comparison of entropy values for letters, words, and digrams between the legal and literary corpora. Entropy is a measure of unpredictability or information content. A higher entropy indicates a more diverse or complex distribution of elements. Key observations include:

- The entropy of *letters* is similar in both corpora, suggesting that the distribution of letters is relatively stable between legal and literary texts.
- For *words*, the entropy is higher in the literary corpus compared to the legal corpus. This indicates a greater diversity of word usage in literary texts, which may reflect more varied vocabulary and stylistic choices.
- The *digrams* (pairs of consecutive words) show the highest entropy in both corpora, with the literary corpus again displaying a slightly higher value. This suggests that literary texts tend to have more complex structures when considering word pairs.

The entropy analysis (Figure 1) provides insight into the linguistic richness and complexity of both corpora, with literary texts generally showing greater variability in word usage and structure compared to legal texts.

To compare the legal and literary corpora, both the Kolmogorov-Smirnov (KS) and Chi-Square tests were applied to the frequency distributions of letters, words, and bigrams. Below are the combined results and qualitative conclusions based on these tests:

- **Letters:** Both the KS and Chi-Square tests indicate that the distributions of letters are not significantly different between the two corpora. This suggests that despite the varying nature of legal and literary texts, the foundational letter frequency distribution remains stable across both domains. The similarity in letter usage reflects commonalities in the core structure of the Romanian language.
- **Words:** For words, both statistical tests similarly show no significant differences in the frequency distributions between the corpora. This implies comparable use of words in both legal and literary texts, largely driven by the frequent use of function words (e.g., prepositions, conjunctions). However, this analysis does not capture differences in content-specific vocabulary, focusing instead on overall word frequency patterns.
- **Bigrams:** The most significant divergence is observed in the bigram analysis. While the Chi-Square test does not indicate a significant difference, the KS test highlights notable discrepancies in the distribution of bigrams between the two corpora. This implies that the legal corpus employs more formulaic and repetitive word pairings, while the literary corpus exhibits greater variation and creativity in word combinations. This difference underscores the formal and structured nature of legal language compared to the more dynamic and flexible use of language in literary texts.
- In summary, while the distributions of letters and words are consistent across both corpora, bigrams show clear differences. The legal corpus is characterized by more structured, repetitive patterns, reflecting its formal nature, whereas the literary corpus demonstrates greater linguistic variety and creativity in word pairings, indicative of more flexible and expressive writing. These results highlight the distinct linguistic profiles of legal and literary texts.

3.3. Conditional Probability Analysis

To further understand the structured nature of legal language, we computed conditional probabilities to analyse the likelihood of specific terms following others in the corpus. This approach allowed us to capture commonly recurring phrase patterns, which are particularly prevalent in legal texts due to the formulaic

and precise language often used in drafting legislation. For instance, phrases such as “art alin” or “in care” demonstrated high conditional probabilities, reflecting their frequent use in legal contexts to maintain consistency and clarity. In contrast, conditional probabilities in the literary corpus showed a broader diversity in word pairings, with less repetition of specific phrases. This analysis underscores the highly structured and predictable patterns characteristic of legal language, as opposed to the more variable structure found in literary texts.

While this study primarily focuses on raw frequency analysis to highlight basic linguistic patterns, we acknowledge the importance of the Term Frequency-Inverse Document Frequency (Tf-Idf) algorithm in identifying terms with high semantic significance. Applying Tf-Idf could enhance the understanding of unique and contextually relevant vocabulary within each corpus, particularly by down-weighting common stopwords and emphasizing distinctive terminology in the legal texts.

The character-level analysis focused on basic frequency counts to capture general distribution patterns within the legal text. We acknowledge that applying Byte Pair Encoding (BPE) could further enhance this analysis by identifying commonly occurring character sequences, which would be particularly beneficial for understanding repeated patterns in formal legal language. BPE is an effective method for grouping frequent character pairs, often leading to a more compressed representation of text and revealing linguistic structures unique to specialized corpora. We recognize the potential of this approach to deepen our insights, especially in the context of standardized legal phrasing, and we will consider incorporating BPE or similar encoding techniques in future studies to build on these preliminary findings.

4. Conclusions

This paper presents the construction and analysis of a representative legal corpus for the Romanian language, focusing on identifying similarities and differences with a general literary corpus. By applying statistical methods such as n-grams, conditional probabilities, entropy, and statistical tests like Kolmogorov-Smirnov and Chi-Square, several key insights were derived.

The legal corpus, composed of over 1 million words drawn from 21 diverse legal texts, offers a comprehensive representation of the formal and standardized nature of legal language. Through the analysis of letter and word frequency distributions, it became evident that there is a high degree of similarity between the legal and literary corpora. This suggests that foundational linguistic structures are consistent across both domains, as confirmed by the Kolmogorov-Smirnov

and Chi-Square tests, which revealed no significant differences in the overall distributions of letters and words.

However, when analysing word bigrams, significant differences between the two corpora emerged. The literary corpus displayed more complex and varied patterns, reflecting the creative and stylistic flexibility inherent in literary writing. In contrast, the legal texts exhibited more rigid and repetitive structures, as expected from formalized legal language. The entropy analysis further highlighted these differences, showing that literary texts possess greater unpredictability and linguistic richness, especially in word usage and bigram combinations, whereas legal texts follow more formulaic patterns.

These findings have important implications for the development of natural language processing (NLP) tools tailored to legal language. The standardized nature of legal texts could enable the creation of specialized algorithms for tasks like automated legal document processing and legal translation. On the other hand, the complex patterns observed in literary texts pose challenges for more generalized language models due to their dynamic and diverse linguistic structure.

In conclusion, although legal and literary corpora share common linguistic foundations, their distinct patterns – particularly in word pairings and complexity – highlight the unique characteristics of each genre. This study provides a strong foundation for future linguistic and NLP research, especially within the legal domain.

References

- Paul Baker. 2006. *Using Corpora in Discourse Analysis*. Continuum.
- Douglas Biber, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.
- cor 2014. Corola descriere generală. <https://corola.racai.ro/> [Accessed: 01.09.2024].
- cor 2024. Legal information institute. <https://www.law.cornell.edu/> [Accessed: 15.09.2024].
- Alexandru Dinu and Adriana Vlad. 2021. Romanian printed language, statistical independence and the type ii statistical error. *International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, p. 120-125.
- Alexandru Dinu and Adriana Vlad. 2022. Revisiting the idea of a representative linguistic corpus. *14th International Conference on Communications (COMM)*, Bucharest, Romania, 2022, p. 1-5.
- Alexandru Dinu, Adriana Vlad, Bogdan Hanu, and Adrian Mitrea. 2019. Revisiting the statistical independence for the printed Romanian language. *Proceedings of the 14th International Conference Linguistic Resources and Tools for Natural Language Processing*, p. 99-113.
- Alexandru Dinu, Adriana Vlad, Bogdan Hanu, and Adrian Mitrea. 2020a. Beginning and end of sentence word digrams for printed Romanian language. *Proceedings of the 15th International Conference Linguistic Resources and Tools for Natural Language Processing*, p. 53–63.

- Alexandru Dinu, Adriana Vlad, Adrian Mitrea, and Bogdan Hanu. 2020b. The statistical independence for words in printed Romanian language. *13th International Conference on Communications (COMM)*, p. 319–324.
- Alexandru Dinu. 2020a. Psycholinguistics, statistics and the unconscious mind. *Lucrare de licență în Psihologie susținută la Universitatea Titu Maiorescu, București.*
- Alexandru Dinu. 2020b. Psycholinguistics, statistics and the unconscious mind. *Conferința Internațională Educație și Creativitate pentru o Societate Bazată pe Cunoaștere – PSIHLOGIE, București, Universitatea Titu Maiorescu, 2020*, p. 190–195.
- eur 2024. Eur-lex. <https://eur-lex.europa.eu/oj/direct-access.html> [Accessed: 01.10.2024].
- Romanian Government. 2016a. Hotărârea 394/2016 pentru aprobarea normelor metodologice de aplicare a prevederilor referitoare la atribuirea contractului sectorial/acordului-cadru din legea nr. 99/2016 privind achizițiile sectoriale. <https://legislatie.just.ro/Public/DetaliiDocument/179284>. Official Gazette of Romania.
- Romanian Government. 2016b. Hotărârea 395/2016 pentru aprobarea normelor metodologice de aplicare a prevederilor referitoare la atribuirea contractului de achiziție publică/acordului-cadru din legea nr. 98/2016 privind achizițiile publice. <https://legislatie.just.ro/Public/DetaliiDocument/179287>. Official Gazette of Romania.
- Romanian Government. 2019. Ordonanța de urgență 57/2019 privind codul administrativ. <https://legislatie.just.ro/Public/DetaliiDocument/215925>. Official Gazette of Romania.
- Bogdan Hanu, Adriana Vlad, Adrian Mitrea, and R. Dragomir. 2016. An analysis of common word digrams in different literary Romanian corpora. *Proc. 11th IEEE International Conference on Communications (COMM2016)*, June 9-10, 2016, Bucharest, p. 43–46.
- Bogdan Hanu, Adriana Vlad, and Adrian Mitrea. 2018. Aspects revealing the orthography and punctuation impact in printed Romanian: A literary corpus-based study. *Proc. 12th IEEE International Conference on Communications (COMM 2018)*, June 14-16, p. 95–100.
- Bogdan Hanu, Adriana Vlad, Alexandru Dinu, and Adrian Mitrea. 2019. Looking along Zipf's Law for the distribution of words beginning and ending sentences in literary printed Romanian corpora. *Proceedings of the 14th International Conference Linguistic Resources and Tools for Natural Language Processing*, p. 51–63.
- Monica Macoveiciuc and Adam Kilgariff. 2010. The RoWaC Corpus and Romanian Word Sketches. In *Multilinguality and Interoperability in Language Processing with Emphasis on Romanian*, Dan Tufiş and Corina Forăscu (eds.) Romanian Academy Publishing House, Bucharest, 2010. ISBN: 978-973-27-1972-5, p 149-166
- Tony McEnery and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction*. Edinburgh University Press.
- Romanian Parliament. 1991. Constituția României din 1991 – republicare. <https://legislatie.just.ro/Public/DetaliiDocument/125327>. Official Gazette of Romania.
- Romanian Parliament. 1996. Legea 8/1996 privind dreptul de autor și drepturile conexe - republicare. <https://legislatie.just.ro/Public/DetaliiDocument/71578>. Official Gazette of Romania.
- Romanian Parliament. 2001. Legea 544/2001 privind liberul acces la informațiile de interes public. <https://legislatie.just.ro/Public/DetaliiDocument/31413>. Official Gazette of Romania.

- Romanian Parliament. 2002. Legea 500/2002 privind finanțele publice. <https://legislatie.just.ro/Public/DetaliiDocument/36627>. Official Gazette of Romania.
- Romanian Parliament. 2003. Codul muncii din 2003 (legea nr. 53 din 24 ianuarie 2003) - republicare. <https://legislatie.just.ro/Public/DetaliiDocument/41982>. Official Gazette of Romania.
- Romanian Parliament. 2004. Legea 554/2004 contenciosului administrativ. <https://legislatie.just.ro/Public/DetaliiDocument/53269>. Official Gazette of Romania.
- Romanian Parliament. 2006a. Codul vamal din 2006 al României. <https://legislatie.just.ro/Public/DetaliiDocument/74729>. Official Gazette of Romania.
- Romanian Parliament. 2006b. Legea 273/2006 privind finanțele publice locale. <https://legislatie.just.ro/Public/DetaliiDocument/72829>. Official Gazette of Romania.
- Romanian Parliament. 2006c. Legea 95/2006 privind reforma în domeniul sănătății - republicare. <https://legislatie.just.ro/Public/DetaliiDocument/70779>. Official Gazette of Romania.
- Romanian Parliament. 2009. Codul penal din 2009 (legea nr. 286/2009). <https://legislatie.just.ro/Public/DetaliiDocument/109855>. Official Gazette of Romania.
- Romanian Parliament. 2010a. Codul de procedură civilă din 2010 - republicare. <https://legislatie.just.ro/Public/DetaliiDocument/120002>. Official Gazette of Romania.
- Romanian Parliament. 2010b. Codul de procedură penală din 2010 (legea nr. 135/2010). <https://legislatie.just.ro/Public/DetaliiDocument/123236>. Official Gazette of Romania.
- Romanian Parliament. 2011. Legea 62/2011 a dialogului social - republicare. <https://legislatie.just.ro/Public/DetaliiDocument/128054>. Official Gazette of Romania.
- Romanian Parliament. 2015a. Codul de procedură fiscală din 2015 (legea nr. 207/2015). <https://legislatie.just.ro/Public/DetaliiDocument/170807>. Official Gazette of Romania.
- Romanian Parliament. 2015b. Codul fiscal din 2015 (legea 227/2015). <https://legislatie.just.ro/Public/DetaliiDocument/170842>. Official Gazette of Romania.
- Romanian Parliament. 2016a. Legea 100/2016 privind concesiunile de lucrări și concesiunile de servicii. <https://legislatie.just.ro/Public/DetaliiDocument/179283>. Official Gazette of Romania.
- Romanian Parliament. 2016b. Legea 101/2016 privind remediile și căile de atac în materie de atribuire a contractelor de achiziție publică, a contractelor sectoriale și a contractelor de concesiune de lucrări și concesiune de servicii, precum și pentru organizarea și funcționarea consiliului național de soluționare a contestațiilor. <https://legislatie.just.ro/Public/DetaliiDocument/179291>. Official Gazette of Romania.
- Romanian Parliament. 2016c. Legea 99/2016 privind achizițiile sectoriale. <https://legislatie.just.ro/Public/DetaliiDocument/179279>. Official Gazette of Romania.
- Marius Sala. 2006. *Aventurile unor cuvinte românești*, volume 1. Univers Enciclopedic.
- Adriana Sferle. 2005. Limbajul juridic și limba comună. https://litere.uvt.ro/litere-old/vechi/documente_pdf/aticole/uniterm/uniterm3_2005/asf_erle.pdf [Accessed: 01.10.2024].
- Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufiş, and Daniel Varga. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias, editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Elena Tognini-Bonelli. 2001. *Corpus Linguistics at Work*. John Benjamins Publishing Company.

Adriana Vlad, Alexandru Dinu, Adrian Mitrea, and Bogdan Hanu. Iunie 2020. Analiză privind grupurile de două cuvinte succesive de început și respectiv de final de frază/propoziție, pe corpus literar de limbă română scrisă cu ortografie și punctuație - faza 1. In *Programul de cercetare: Resurse și tehnologii pentru limba română în context multilingv standardizat*. Academia Română, Secția de Știința și Tehnologia Informației, Institutul de Cercetări pentru Inteligență Artificială.

Adriana Vlad, Alexandru Dinu, Adrian Mitrea, and Bogdan Hanu. Noiembrie 2020. Analiză privind grupurile de două cuvinte succesive de început și respectiv de final de frază/propoziție, pe corpus literar de limbă română scrisă cu ortografie și punctuație - faza 2. In *Programul de cercetare: Resurse și tehnologii pentru limba română în context multilingv standardizat*. Academia Română, Secția de Știința și Tehnologia Informației, Institutul de Cercetări pentru Inteligență Artificială.

ROMANIA IN A MULTILINGUAL CORPUS OF EUROPEAN POETRY

SILVIE CINKOVÁ¹, PETR PLECHÁČ²

¹ *Charles University, Faculty of Mathematics and Physics and Institute of the
Czech Literature of the Czech Academy of Sciences*
cinkova@ufal.mff.cuni.cz

² *Institute of the Czech Literature of the Czech Academy of Sciences*
plechac@ucl.cas.cz

Abstract

This paper presents a pilot research on Romanian geo-entities in European Poetry, performed on the PoeTree corpus. PoeTree is a harmonized multilingual corpus of European poetry, which currently contains documents in Czech, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian, and Spanish, mostly dating back to the 19th century. It promises to provide manually edited entity linking. That is still in progress and the linked data set is not yet available, but by courtesy of PoeTree developers, we could identify poems with at least one mention of a Romanian geo-entity.

Keywords: distant reading poetry entity linking corpus computational literary studies

1. Introduction

How is Romania reflected in the European poetry? We have performed a preliminary research on PoeTree (Plecháč and Kolrá, 2015), a harmonized corpus of European poetry, which currently features ten languages: Czech, German, English, Spanish, French, Hungarian, Italian, Portuguese, Russian, and Slovenian. It has a convenient REST API, and the next release is going to contain linked WikiData entities. From the ongoing annotation project, we have extracted a list of Romanian WikiData entities currently represented in PoeTree, along with the IDs of the poems where they occurred. In this pilot study, we exploit this improvised data set.

The goal of the study was to explore one possible workflow and assess how useful the WikiData linking in PoeTree could be for a literary scholar or historian with basic data-mining skills using this particular workflow. Following this setup, we have even sketched out some conclusions a scholar could draw, although we are well aware of the current data set being incomplete and our knowledge of

Romanian literature and history being too superficial to formulate and answer serious research questions. Nevertheless, we demonstrate how this workflow, once applied to the complete data set, can be used in data-driven research in (literary) history.

2. Data and Annotation

2.1. Cleaning and Harmonization

PoeTree is a multilingual data set of European poetry. It combines several monolingual poetry corpora:

- The Corpus of Czech Verse (Plecháč and Kolrá, 2015)
- German Poetry Corpus (Bobenhausen and Hammerich, 2015; Haider, 2021; Haider, 2024; Lohmeier et al., 2012)
- Corpus of Spanish Golden Age Sonnets (Navarro et al., 2016)
- Diachronic Spanish Sonnet Corpus (Navarro et al., 2016)
- Corpus Malherbae (Delente and Renault, 2023)
- Biblioteca Italiana (Quondam, 2021)
- Poemas (Mittman et al., 2019)
- Corpus of Russian Poetry (Grishina et al., 2009)
- ELTE Poetry Corpus (Horváth et al., 2022)
- A Gutenberg Poetry Corpus (Parrish, 2018)
- A corpus of Slovenian poetry¹.

These monolingual corpora were harmonized to a common format, and the data were cleaned by de-duplicating the poems and removing poems that were not (mainly) written in the declared language. Duplicates were not removed but marked. Then the poems were tokenized, lemmatized and morpho-syntactically tagged with the Universal Dependencies formalism (Nivre et al., 2020), mostly using the UDPipe2.0 parser (Straka et al., 2016).

3. Named-Entity Linking

The data were also enriched by VIAF IDs and WikiData entities of the authors whenever possible. In a next step, named entities were recognized in the poems (using NameTag2 (Straková et al., 2019) for Czech, SpaCy for English, and GPT 4 Turbo for the other languages, the selection being based on an evaluation on manually annotated PoeTree samples (Plecháč et al., 2024)).

¹ Built by Urška Perenič's project Wikiprojekt Slovenska Leposlovna Klasika, not on public display outside PoeTree (Perenič, 2024).

The recognized named entities were automatically linked to WikiData. This annotation will allow for cross-lingual content analysis: for instance, no matter whether a place is spelled *Roma*, *Rome*, *Rom*, or *Řim* in different languages, it is linked to a common WikiData ID. Since freely available tools for automated entity linking gave disappointing results, each entity was simply linked to the first matching item offered by the WikiData API. At the time of writing this paper, human annotators are manually checking and revising the linking.

To illustrate the extent of the work: the English and the German part of PoeTree contain 11,000 and 12,000 recognized named entities, respectively. Since the algorithm simply picked the first WikiData item matching the entity string, many links are wrong and must be manually corrected. The annotators received a list of recognized entities and the matched WikiData items. They can see the entity string and the assigned WikiData ID in a browser window. Their task is to revise each by viewing the WikiData item associated with the given ID, checking the WikiData database in a separate frame. They can also evoke a pop-up window with poem lines where the given entity string has occurred. They are encouraged to perform additional research to optimize the linking.

Sometimes links are matched to a wrong WikiData item. This happens especially when several WikiData items are represented by the same string. For instance, references to places referring to the Bible or to Classical Antiquity are often mis-linked to their namesakes in the United States (e.g. *Zion*). Such cases are easily disambiguated by the context, but the disambiguation becomes tedious with ambiguous geo-names within the focused region like *York*, or *Falkenstein*, *Rabenstein*, or even person names like *Frederi(c)k* and *Christian* contextualized as Danish kings (who kept succeeding each other in an almost perfect order from the 15th century onward). Sometimes, a place can be only disambiguated because the poem associates it with a historical event, typically a battle, or a person with an ambiguous name turns out to be a well-known historical figure (e. g., the Habsburg Empress Maria Theresa often occurs as *Theresia*), or even someone known only from the author's biography. Therefore, the annotators are instructed to put a reasonable effort into assigning links as specific as possible, the allocated time and budget permitting.

The linking fails when either the entity has no corresponding WikiData item at all, or when the string deviates from the form listed in WikiData due to spelling differences or lemmatization errors. Spelling deviations are frequent in all languages of PoeTree, since most included poems stem from the 19th century or earlier, while the NLP tools are trained on modern prose. Therefore, many entities are lemmatized incorrectly. The lemmatization errors even tend to follow a pattern. For instance, when a name ends with *-s* in German, the lemmatizer would consider the *-s* ending part of the word stem and keep it. It has learned from the modern prose that geo-names in genitive occur rarely, whereas stems ending

with *-s* are rather common. Then the string would not be matched against WikiData at all, or it would be matched to a WikiData item that would contain a precise match of this string. One example demonstrates both the mis-recognized genitive and outdated German spelling: *Gothland* in the context *Gothlands Strand* (*Gotland's coast*) would be linked to *Gothlandstraße* (Q106259706, street in Lübeck, Schleswig-Holstein, Germany) instead to *Gotland* (Q1027830, island and historical province in Sweden). The same problem applies to derived words, for instance adjectives of origin: *Leipziger* (e.g., in *Leipziger Messe*, the *fair of Leipzig*) would be mapped to *Leipziger* (inhabitant of Leipzig) rather than to *Leipzig*.

When no matching WikiData item exists but the string is easily tracked in other resources (e.g., Google maps), the annotator tries to find a WikiData entity located in its proximity. For instance, a neighbourhood can be represented by the corresponding town, a hill can be represented by a mountain nearby, or a historical region can be represented by a modern region with which it has the largest geographical overlap. Hence the linked names need not be true synonyms.

At this stage, the annotation only concerns named entities automatically recognized as geo-names. When a recognized entity turns out to be a person name, the annotators are instructed to provide a WikiData link anyway and mark it as a person name. Therefore, some personal names are going to occur in the next data release, without any claim of completeness. The next PoeTree release including the geo-names WikiData entity linking for selected languages is scheduled for 2025.

4. Distribution of Romanian Entities in PoeTree

4.1. Frequency

Thanks to the steady progress of the manual WikiData entity linking, we could perform a tentative exploration of entities related to Romania that occur in different languages in PoeTree. The current set contains sixty-seven unique geo-entities referring to Romania (Cf. Fig. 1). The simplest perspective is the distribution of entities in languages and in documents. Currently, most unique Romanian entities occur in Hungarian, Czech, Russian, and German poems (See Tab. 1).

ROMANIA IN A MULTILINGUAL CORPUS OF EUROPEAN POETRY



Figure 1: Frequency of all Romanian geo-entities in collections

Another distribution aspect is in how many different poems in the given language an entity occurs and how many times; the more poems with recurrent entities there are in a given language, the more likely it is that Romania arouses interest in poets writing in the given language. This view is plotted in Fig. 1). The y axis shows how many times an entity occurred in one poem. The x axis shows in how many poems this was the case. The plot is divided in sub-graphs of the individual languages. So, for instance, we can trace one Hungarian poem where an entity occurred sixty-two times. It happened to be *Timișoara*. At the same time, only one Romanian entity occurred in the (so-far wiki-linked) French, Italian, and Slovenian collections in PoeTree, respectively. In Slovenian, it was *Dniester* that made it to four different poems, in French as well as in Italian it was *Dacia*.

5. TF*IDF

The most frequently mentioned entities across languages are the *Carpathian Mountains* (162), *Transylvania* (128), *Timișoara* (93), *Oradea* (57), *Dniester* (40), and *Izmail* (22). However, frequency does not say how typical of a language collection an entity is, among other named entities, or among all content words.

A more appropriate metric would be TF*IDF. The abbreviation stands for *Term Document Frequency multiplied by the Inverted Document Frequency*. It measures the importance of individual words for individual documents in a document collection. The collection of texts is here all PoeTree poems containing an entity referring to Romania. Each word has one Term Document Frequency (TF) value for each document in the collection. The frequency count is relative; that is, the values are between 0 and 1. When a word does not occur in a document at all, its TF equals 0. Should a document consist only of that one word (repeated or single), the TF of that word for that document would be 1. The Inverse Document Frequency (IDF) counts in how many documents of the collection that word occurred (no matter how many times). More precisely, it is the total number of documents in the entire collection divided by the number of documents that contain the given word, and this number is transformed with a logarithm. The TF*IDF score of a word for a document is the higher, the more frequent the given word is in the given document and the fewer documents this word occurs in.

To measure TF*IDF in a collection, all texts must be written in the same language. To bypass PoeTree’s multilinguality, we had GPT-4.0 translate the entire collection of poems referring to Romania. We used the following prompt: “You are an expert translator. Translate this text to plain modern English prose”. The temperature was set to 0 to make the translation as deterministic as possible. English originals were translated in the same way. A random manual check suggested a very good translation quality.

We tagged and lemmatized the English output, along with recognizing named entities, with UDPipe (Straka et al., 2016; Wijffels, 2023), so the English data set was prepared in exact the same way as the translations.

In the first experiment, we kept only named entities in the texts and computed their TF*IDF. As expected, not many Romanian entities were among the top ten TF*IDF values, except in the Hungarian (*Transylvania*) and Russian (*Kagul*, *Kishinev*, *Rymnik*) poetry.

Transylvania is very frequent in all languages, which decreases its TF*IDF for all documents; therefore, it must have been extraordinarily present in the Hungarian collection to compensate for the high frequency in all other documents and rank among the top ten (See Fig. 3a). Also, *Várad*, which originally is a Hungarian word for *castle* or *fortification*, could be referring to *Oradea*.

The Russian collection unveils a possible focus on battles in the historical Romanian territory, probably those against Turks, since the Romanian top TF*IDF-ranked entities are accompanied by *Istanbul* and *Saracens*. Moreover, *Suvorov* refers to a Russian military commander who won a battle against the Turks at *Rymnik* (1799).

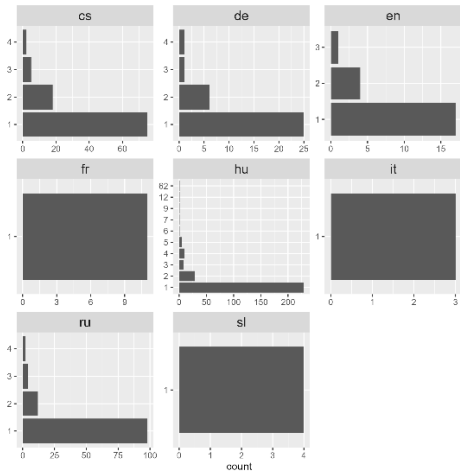


Figure 2: Distribution of entities in different poems

language	NE
hu	40
cs	19
ru	17
de	13
en	7
fr	6
sl	3
it	1
es	0
pt	0

Table 1: Distribution of unique Romanian named entities in PoeTree’s languages

The German collection features *Romania* and *Bessarabia* (*Bessa*) among the top ten TF*IDF, but no connection to the other named entities is apparent. The only entity in the Slovenian collection that potentially relates to Romania is *Danube*, but the other top-ten-TF*IDF ranked named entities do not hint to its specifically Romanian locations.



(a) Top ten most **typical** entities
(TF*IDF)

(b) Top ten most **frequent** entities
(TF)

Figure 3: Top-ten named entities on lemmatized English translations, in each language corpus separately

6. Term Frequencies (TF) of Named Entities in the Individual Language Collections

The top ten term frequencies (TF) of named entities in the collections (See Fig. 3b) reveal Czech references to *Danube* and the *Carpathian mountains*, neither being specific to Romania. The TF view of the German, Hungarian, and Russian collections is similar to the TF*IDF perspective. The English collection has *Dacia* among geographic references wide-spread among Spain and Pacific. The Italian collection's TF view captures *Arad* and the Slovenian also *Dniester* beside the *Danube* and *Carpathians*.

7. TF*IDF of All Lemmatized Content Words

The last facet obtained from the English translations presents the top ten TF*IDF content words. There were many ties, and therefore some language subgraphs plot in Fig. 4 contain more than ten words. In this perspective, there is no trace of a specifically Czech, German, English, Slovenian, or Italian reference to Romania. Again, the Hungarian and Russian collection preserve some, possibly both associating Romania with Turks (*Pasha* and *Transylvania* in Hungarian and *Suvorov* in Russian).



Figure 4: TF*IDF of content words in lemmatized English translations, grouped by languages, top ten

8. Discussion

This explorative study admittedly suffers from many limitations: the automatic procedures display apparent flaws: so, some entities get lost in the automatic translation; the tagger does not recognize all named entities correctly; the lemmatizer sometimes generates incorrect lemmas. In spite of these limitations, we could quickly gain orientation in Romanian references in the European poetry, as it is represented by PoeTree, and generate a preliminary insight that the Hungarian and Russian poetry referring to Romania was decidedly martial, but apparently concentrating on defence against Turkish raids on Europe rather than on domination over the Romanian population. With linked entities in place and incorporated into the automatic translation to modern English prose, the results are expected to become much more informative.

Acknowledgements

This work has been supported by the Czech Science Foundation grant No. GA23-07727S and has also been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

- Klemens Bobenhausen and Benjamin Hammerich. 2015. Métrique littéraire, métrique linguistique et métrique algorithmique de l’allemand mises en jeu dans le programme Metricalizer2. *Langages*, 199(3):67–88.
- Éliane Delente and Richard Renault. 2021. Projet Anamètre : présentation, limites et avancées. In Anne-Sophie Bories, Gérald Purnelle, and Hugues Marchal, editors, *Plotting Poetry, On Mechanically-Enhanced Reading*, pages 73–92, Liège. Presses universitaires de Liège.
- Elena Pavlovna Grishina, Kirill Michailovich Korchagin, Vladimir Plungian, and Dmitry Sichinava. 2009. Poeticheskii korpus v ramkah NKRIA: obschaia struktura i perspektivy ispolzovania. In *Natsionalnii korpus russkogo iazyka: 2006–2008. Novye rezultaty i perspektivy*, pages 71–113. Nestor-Istoria.
- Thomas Haider. 2021. Metrical Tagging in the Wild: Building and Annotating Poetry Corpora with Rhythmic Features. In Merlo Paola, Tiedemann Jorg, and Tsarfaty Reut (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3715–3725, Online. Association for Computational Linguistics.
- Thomas Haider. 2024. A Large Annotated Reference Corpus of New High German Poetry. In Calzolari Nicoletta, Kan Min-Yen, Hoste Veronique, Lenci Alessandro, Sakti Sakriani, and Xue Nianwen (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 677–683, Torino, Italia. ELRA and ICCL.

- Péter Horváth, Péter Kundráth, Balázs Indig, Zsófia Fellegi, Eszter Szilávi, Tímea Borbála Bajzát, Zsófia Sárközi-Lindner, Bence Vida, Aslihan Karabulut, Mária Timári, and Gábor Palkó. 2022. ELTE Poetry Corpus: A Machine Annotated Database of Canonical Hungarian Poetry. In Calzolari Nicoletta, Béchet Frédéric, Blache a Philippe, Choukri Khalid, Cieri Christopher, Declerck Thierry, Goggi Sara, Isahara Hitoshi, Maegaard Bente, Mariani Joseph, Mazo Hélène, Odijk Jan, and Piperidis Stelios (eds.). *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3471–3478, Marseille, France. European Language Resources Association.
- Felix Lohmeier, Kathleen Smith, Sibylle Söring, and Ubbo Veenster. 2012. The TextGrid Repository: Supporting the Data Curation Needs of Humanities Researchers.
- A Mittman, Alckmar Luiz dos Santos, and Paulo Henrique Pergher. 2019. What Rhythmic Signature Says about Poetic Corpora. In *Quantitative Approaches to Versification*, pages 153–172, Prague, Czech Republic. ICL CAS.
- Borja Navarro, María Ribes Lafoz, and Noelia Sánchez. 2016. Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. In Calzolari Nicoletta, Choukri Khalid, Declerck Thierry, Goggi Sara, Grobelnik Marko, Maegaard Bente, Mariani Joseph, Mazo Helene, Moreno Asuncion, Odijk Jan, and Piperidis Stelios (eds.). *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4360–4364, Portorož, Slovenia. European Language Resources Association (ELRA).
- Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.
- Allison Parrish. 2018. A Gutenberg Poetry Corpus. Online. GitHub.
- Urška Perenič. 2024. Slovenska leposlovna klasika. https://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika.
- Petr Plecháč, Silvie Cinková, and Robert Kolár. 2024a. Rozpoznávání Pojmenovaných Entit v Básnických Textech. In *Book of Abstracts*, pages 102–103, Prague, Czech Republic. Charles University, Faculty of Arts.
- Petr Plecháč and Robert Kolár. 2015. The Corpus of Czech Verse. *Studia Metrica et Poetica*, 2(1):107–118.
- Petr Plecháč, Robert Kolár, Silvie Cinková, Artjoms Šeļa, Mirella De Sisto, Lara Nugues, Thomas Haider, Neža Kočnik, Benjamin Nagy, Éliane Delente, Richard Renault, Klemens Bobenhausen, Benjamin Hammerich, Adiel Mittmann, Gábor Palkó, Péter Horváth, Borja Navarro Colorado, Pablo Ruiz Fabo, Helena Bermúdez Sabel, et al. 2024b. *PoeTree 0.0.2. Poetry Treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian, Slovenian and Spanish*. Online. Zenodo.
- Amedeo Quondam. 2021. Memorie per una storia dell'italianistica digitale: Biblioteca italiana. *Griseldaonline*, 20(2):137–147. Online.
- Pablo Ruiz Fabo, Helena Bermúdez Sabel, Clara Martínez Cantón, and Elena González-Blanco. 2020. The Diachronic Spanish Sonnet Corpus: TEI and linked open data encoding, data distribution, and metrical findings. *Digital Scholarship in the Humanities*, 36(Supplement 1):i68–i80.

- Milan Straka, 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'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).
- Jana Straková, Milan Straka, and Jan Hajic. 2019. Neural Architectures for Nested NER through Linearization. In Korhonen Anna, Traum David, and Màrquez Lluís (eds.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.
- Jan Wijffels. 2023a. udpipe: Tokenization, Parts of Speech Tagging, Lemmatization and Dependency Parsing with the 'UDPipe' 'NLP' Toolkit. R package version 0.8.11.

HUMAN VERSUS AI-REWRITTEN ACADEMIC WRITING: A CORPUS-BASED COMPARATIVE ANALYSIS OF PHRASEOLOGICAL COMPLEXITY IN ESSAYS AND SCIENTIFIC ARTICLES

MĂDĂLINA CHITEZ¹, ROXANA ROGOBETE¹,
ANA-MARIA BUCUR², ANDREEA DINCA¹, IOANA-SIMINA FRÎNCU³

¹*West University of Timișoara / CODHUS, Timișoara, Romania*
{madalina.chitez, roxana.rogobete,
andreea.dinca}@e-uvt.ro

²*niversity of Bucharest, ISDS, Romania*
ana-maria.bucur@drd.unibuc.ro

³*West University of Timișoara / ICAM, Timișoara, Romania*
ioana.frincu@e-uvt.ro

Abstract

As AI-generated academic writing becomes more common, understanding how it compares to human-authored texts, particularly in Romanian academic writing, is essential. This study compares the original academic texts from two Romanian writing corpora, ROGER (Corpus of Romanian Academic Genres; Chitez et al., 2021), which includes student essays, and EXPRES (Corpus of Expert Writing in Romanian and English; Chitez et al., 2022), which features scientific articles, against the same texts rewritten by ChatGPT. By analysing key linguistic features such as adjectival and adverbial modifiers, direct objects, as well as specific corpus-based multi-word expressions, this research explores the phraseological complexity and sophistication (Paquot, 2019) in both AI-rewritten and human-authored texts. This study aims to investigate how, and if, ChatGPT improves the quality of academic texts, both student essays and expert scientific articles, by increasing their phraseological complexity. The results show that while ChatGPT can generate more structured and sophisticated phrases, it tends to reduce overall diversity, often leading to more formulaic constructions. For novice writers, such as those represented in the ROGER corpus, the ability of ChatGPT to refine grammar and increase phraseological sophistication may be useful for improving clarity and coherence in their academic essays. However, the reduced diversity might limit the development of more varied and creative expression. For expert writers, as represented in the EXPRES corpus, the AI-generated texts replicate some of the sophisticated phraseological patterns found in scientific articles, but they often lack the nuanced and complex structures that expert-level academic writing demands. These findings highlight both the

strengths and limitations of large language models (LLMs) like ChatGPT in supporting different levels of academic writing in the Romanian language.

Keywords: Romanian Academic Writing Corpora, student writing versus ChatGPT writing, expert writing versus ChatGPT writing, phraseological complexity in academic writing, phraseological complexity in human versus AI-generated academic writing, essay AI-rewriting, scientific article AI-rewriting

1. Introduction

In the AI-driven landscape of academic writing support, it becomes essential to evaluate the linguistic features of texts refined or generated by large language models (LLMs) like ChatGPT. Phraseological complexity is one of the features which signals the level of proficiency in academic text writing (Vandeweerd et al., 2021). It reflects the writer's ability to use sophisticated multi-word expressions and collocations, which increases with higher proficiency levels, as shown in studies on L2 learners (Paquot, 2018; Paquot, 2019). This study builds on these foundations by comparing the phraseological sophistication of AI-generated texts with human-authored texts from novice and expert academic writers, using academic writing (AW) corpora specific to the Romanian language context. The two corpora are ROGER (Corpus of Romanian Academic Genres) (Chitez et al., 2021; Chitez, 2022) and EXPRES (Corpus of Expert Writing in Romanian and English) (Chitez et al., 2022). Both corpora are available online for free queries via dedicated corpus-based academic writing platforms¹ (details in sections 2.2 and 3.1).

Our aim is to investigate how, and if, ChatGPT improves the quality of academic texts, both student essays and expert scientific articles, by increasing their phraseological complexity. By focusing on key linguistic features such as adjectival and adverbial modifiers and multi-word expressions, we seek to explore how effectively AI replicates or diverges from human phraseological patterns in academic contexts. These insights will inform ongoing discussions on the role of AI in academic writing and its potential to support or complement human creativity and expertise.

2. Previous Research

2.1. NLP Academic Writing studies in Romania

Quite few studies have dealt with the analysis of academic writing in Romania e.g. (Ene and Mitrea, 2013; Baniceru and Tucan, 2018), and even fewer have used

¹ <https://roger-corpus.org/> and <https://expres-corpus.org/>

corpus linguistics to conduct rigorous investigations. ROGER, a novice writing corpus, and its multiple metadata (Chitez et al., 2021), often set in contrast with the corpus EXPRES (Chitez et al., 2022b), representing expert academic writing in the Romanian context, has generated several significant research findings. A study by Muresan et al. (2022) addresses patterns of co-occurring words and collocations leading to the identification of multi-word linguistic units specific to various disciplines. Furthermore, a chapter (Bercuci and Chitez, 2023) was dedicated to the “Romanian writing style,” revealed through various linguistic features in the English writing of a Romanian native speaker. ROGER has also been used to investigate authorial voice in Bachelor theses (Doroholschi et al. 2025, in press).

In point of generated linguistic resources based on AW corpora, Bucur et al. (2022) compiled the first Romanian Academic Word List (Ro-AWL). Ro-AWL was developed using a combination of computational linguistics methods, corpus analysis, and approaches from L2 academic writing. The authors used two main corpora for the compilation of Ro-AWL: the ROMBAC corpus, a general language corpus, and the EXPRES corpus, which contains expert academic writing from four disciplines: Linguistics, Economics, Information Technology, and Political Science. The Ro-AWL was created following methodologies used for English academic vocabulary lists, particularly focusing on frequency, dispersion, and disciplinary measures to identify words commonly used across academic texts but less frequent in general language. The resulting list contains 673 lemmas distributed across key parts of speech, such as nouns, verbs, adjectives, and adverbs. Ro-AWL holds pedagogical significance, offering a useful resource for teaching academic writing in Romanian, while also serving as a foundation for Natural Language Processing (NLP) tasks, such as text classification and language modelling. In a follow-up study, Chitez et al. (2024) introduced the Romanian Phrasal Academic Lexicon (ROPAL), a resource aimed at supporting academic writing in Romanian by identifying and categorizing frequently used academic phrases used in Romanian academic texts. Academic phrases, such as collocations and multi-word units, are essential for structuring arguments, comparing perspectives, and signalling transitions in discourse. In ROPAL, these phrases are classified based on rhetorical functions, drawing on frameworks like the Oxford Phrasal Academic Lexicon (OPAL). This resource can serve not only as a pedagogical tool for improving academic writing but also as a foundation for computational models that enhance automated academic writing assessments and the development of AI-based tools for academic text generation in Romanian.

Studies have been conducted on the assessment of academic writing phraseological complexity as well. A study by Dinca et al. (2024) focused on three main phraseological components: adjectival modifiers, adverbial modifiers, and direct objects. Statistical analysis reveals that novice L1 Romanian writers use

more diverse adjectival modifiers (mean = 6.05) than the Romanian writers writing in L2 English (mean = 5.29), with a significant p-value of $p < 0.001$. Conversely, novice L2 writers exhibit more diversity in direct objects (mean = 4.12) compared to L1 writers (mean = 3.70, $p < 0.001$). Regarding sophistication, novice L1 writers outperform L2 writers in the use of adjectival modifiers (PMI = 3.80 vs. 3.07, $p < 0.001$) and direct objects (PMI = 3.39 vs. 2.36, $p < 0.001$). The results of this study have important implications for teaching academic writing, particularly for students transitioning from L1 Romanian to L2 English. The findings indicate that Romanian novice writers tend to transfer simpler, less diverse, and less sophisticated phraseological structures from Romanian to English, especially in the use of adjectival modifiers and direct objects. This suggests that targeted teaching interventions should focus on helping students expand their range of phraseological constructions in L2 English.

2.2. Existing Tools

For under-resourced languages, such as Romanian, the existence of any corpus-based tools or linguistic tools that can support L2 English language learners or L1 Romanian academic writing learners can have a significant impact on the teaching and research community. For example, the sophisticated and educationally relevant NLP platform ReaderBench (Dascalu et al., 2017), offers free access to an automated linguistic evaluation platform that holds significant applicability across various fields, such as pedagogy, media, and automatic language processing. At the same time, the integrative platform RELATE (Romanian Portal of Language Technologies, proposed by RACAI (Research Institute for Artificial Intelligence “Mihai Drăgănescu” of the Romanian Academy), includes features such as: the Romanian Reference Corpus (CoRoLa, Mititelu et al., 2014) or RoWordNet (Tufiş et al., 2006), NER etc. Unfortunately, some of them are not accessible to the general user. None of these includes samples of learner language or searchable batches of academic texts. Until now, the only available free-access academic writing resources (L1 Romanian and L2 English, both novice and expert, writing) have been developed at the CODHUS research centre at the West University of Timisoara, Romania. For example, the ROGER and EXPRES corpora (more information in section 3.1) are freely available for research and pedagogical purposes and can be accessed via dedicated query platforms: ROGER platform (Chitez et al., 2021) and EXPRES platform (Chitez et al., 2022). ROGER has also been included in the CLARIN database (Chitez, 2022). Below we will demonstrate the versatility of these resources and how it has been used for the current study. The approach can be replicated for similar contrastive NLP studies on academic writing in the Romanian context.

2.3. Corpus Based Didactic Applications

The teaching support applicability potential of the ROGER and EXPRES corpora can be exemplified as follows:

- Interference L1 versus L2: For example, Oravitan et al. (2022), identified cross-contamination phenomena between L1 Romanian and L2 English in academic phraseology use.
- DDL L2 error correction: Students can access the ROGER platform to check how a certain word or phrase is used in context. Then they might verify the same with the expert corpus ROGER, thus identifying learner versus expert linguistic use.
- Discipline-specific L2 academic writing patterns: Like the previous case, students can access the ROGER platform to extract discipline-specific word-in-context occurrences in novice versus expert writing. This option can be exploited by teachers as well, who can teach discipline specific vocabulary in both L1 and L2 by considering identified error patterns and differences between learners and experts.
- Specific NLP analyses: Platform administrators can access the ROGER and the EXPRES corpora, filter academic genres, disciplines or level of study and download datasets which can be easily processed using NLP techniques to extract patterns across genres, disciplines or by level of study (e.g., BA versus MA studies). Linguistic complexity assessment can be performed similarly.

3. Data and Methods in This Study

3.1. ROGER and EXPRES

The ROGER corpus – Corpus of Romanian Academic Genres (Chitez et al., 2021; Chitez, 2022) is a specialized academic writing resource that focuses exclusively on texts produced by Romanian university students. It includes a variety of academic genres, such as essays, research papers, and BA theses, across disciplines like Economics, Humanities, Computer Science, and Political Science. ROGER was compiled between 2019 and 2022 and it consists of 913 texts in Romanian and 1,139 texts in English as a foreign language. The primary goal of the ROGER corpus is to analyse and document the academic writing practices of Romanian students, particularly in L1 Romanian and L2 English contexts. This corpus provides valuable insights into how students develop their academic writing skills in both their native language and as learners of English. It serves as a digital resource for understanding the phraseological complexity and challenges that students face in their transition from novice to more advanced academic writers.

The EXPRES corpus – Corpus of Expert Writing in Romanian and English (Chitez et al., 2022) is a bilingual academic writing corpus containing expert-level research articles written by scholars from Romania, as well as native English-speaking researchers. The corpus focuses on a similar set as ROGER of key academic disciplines: Linguistics, Economics, Political Science, and Information Technology. EXPRES was compiled between 2021 and 2022, and it consists of 800 texts in Romanian, 780 texts in English as a foreign language, and 703 texts in English L1. As it includes both L1 Romanian and L2 English texts written by Romanian scholars as well as native English texts, it provides a rich dataset for comparative studies of phraseological complexity, sophistication, and academic writing practices between expert Romanian and native English writers. The EXPRES corpus is particularly valuable for analysing how Romanian researchers navigate academic writing in both their native and second languages.

3.2. ROGER-AI and EXPRES-AI

In order to capture the main transformation processes involved in widely used LLM models such as ChatGPT, we created pilot corpora, based on subsets of ROGER and EXPRES, which were then uploaded in ChatGPT with the prompt “please rewrite this text”. The AI corpora are rewritten texts as follows: ROGER-AI, 30 essays, and EXPRES-AI, 15 articles.

Table 1: Corpus data description

Corpora	No doc	No words	Unique words	Mean words/doc
ROGER	30	39,155	9,653	1,305
ROGER_AI	30	18,311	5,442	610
Expert_L1_RO	15	18,432	5,766	1,228
Expert_L1_RO_AI	15	13,923	4,529	928
Full_Expert_L1_RO	748	3,480,420	225,064	4,653

3.3. Methods

The study employs phraseological complexity metrics, focusing specifically on three dependency pairs: adjectival modifiers, adverbial modifiers, and direct objects. The corpora were processed using Python, utilizing tools such as spaCy for tasks including part-of-speech tagging, dependency parsing, and lemmatization. To evaluate phraseological diversity and sophistication, methodologies from Paquot (2019) and Vandeweerd et al. (2021) were implemented. Statistical tests, like the Kruskal-Wallis and Dunn’s post-hoc with

Bonferroni adjustments, were used to determine notable differences in phraseological complexity between novice and expert writers.

4. Analysis and Results

As mentioned, we used a subset of ROGER, formed of 30 essays written by university students, and a subset of EXPRES, composed of 15 research articles, and compared this set with the texts rewritten by ChatGPT (ROGER_AI and EXPRES_AI).

The analysis of the initial corpus data compared to the AI-generated corpus reveals notable differences regarding general features. In terms of corpus size, the ROGER corpus comprises 30 documents with a total of 39,155 words and 9,653 unique words, while the AI-generated ROGER_AI corpus contains significantly fewer words, totalling 18,311, and only 5,442 unique words. The average word count per document is also lower in the AI corpus (610 words per document), compared to the initial corpus (1,305 words per document). This reduction in both total word count and unique words suggests that the AI-generated texts are more concise, potentially reflecting the AI's tendency to simplify or condense content. A similar trend is observed in the expert-level corpus, where the EXPRES corpus has 18,432 words and 5,766 unique words, while the AI-generated EXPRES_AI corpus contains 13,923 words and 4,529 unique words. This reduction in length and variety in both novice and expert texts produced by AI suggests that while AI can replicate academic writing patterns, it tends to produce more streamlined, less diverse output.

4.1. Phraseological Diversity: ROGER-student versus AI Texts

When examining phraseological diversity, all the differences in *amod*, *advmod* and *dobj* are statistically significant, suggesting that significant differences emerge between the original, student-written and AI-generated texts.

Table 2: Phraseological diversity ROGER versus AI texts

	ROGER		ROGER_AI		
	Mean	Standard Deviation	Mean	Standard Deviation	
Amod	5.87	0.36	7.08	0.35	p=0.01
Advmod	2.71	0.2	1.78	0.11	p<0.001
Dobj	3.99	0.14	4.76	0.19	p<0.004

In the students' texts, adjectival modifiers (Amod) and direct objects (Dobj) display higher diversity compared to the AI-generated texts, with statistically significant differences ($p=0.01$ for Amod and $p=0.004$ for Dobj). Conversely, adverbial modifiers (Advmod) are less diverse in the AI-generated texts, as reflected by a statistically significant decrease in diversity ($p<0.001$). This suggests that while the AI successfully replicates some phrase structures, such as verb-noun combinations, it may simplify others, leading to reduced overall diversity.

4.2. Phraseological Sophistication: ROGER-student versus AI Texts

In terms of phraseological sophistication, the AI-generated ROGER_AI corpus outperforms the original ROGER corpus only in direct object sophistication ($p=0.007$), suggesting that the AI-rewritten texts employ more sophisticated verb-noun combinations.

Table 3: Phraseological sophistication ROGER versus AI texts

	ROGER		ROGER_AI		
	Mean	Standard Deviation	Mean	Standard Deviation	
Amod	3.83	0.13	4.20	0.14	$p=0.65$
Advmod	4.55	0.35	4.37	0.39	$p=0.11$
Dobj	3.50	0.12	4.05	0.17	$p=0.007$

However, no significant differences are observed in adjectival and adverbial sophistication between the original and AI-generated novice texts. This indicates that the AI's enhancement in sophistication is more focused on specific dependency structures, particularly direct objects, while other elements remain unchanged.

4.3. Phraseological Diversity: EXPRES-expert versus AI Texts

The expert-level corpus shows no statistically significant differences in diversity between the original and AI-generated versions, although the AI texts display a slight increase in adjectival and direct object diversity. However, these results are not statistically significant due to the small sample size, indicating that the AI can replicate the complexity of expert-level writing to a certain extent, but without significantly altering the diversity of phraseological structures.

Table 4: Phraseological diversity EXPRES versus AI texts

	EXPRES		EXPRES_AI	
	Mean	Standard Deviation	Mean	Standard Deviation
Amod	7.32	0.58	8.02	0.55
Advmod	1.84	0.21	1.64	0.21
Dobj	3.14	0.25	3.58	0.27

4.4. Phraseological Sophistication: EXPRES-expert versus AI Texts

Similarly, in the expert-level texts, although the results are not statistically significant, the AI-generated texts show a slight improvement in adverbial sophistication, suggesting that AI has the potential to match or even slightly improve the phraseological complexity of expert writing.

Table 5: Phraseological sophistication EXPRES versus AI texts

	EXPRES		EXPRES_AI	
	Mean	Standard Deviation	Mean	Standard Deviation
Amod	4.14	0.26	4.23	0.29
Advmod	4.21	0.65	4.90	0.34
Dobj	3.18	0.27	3.52	0.21

4.5. Corpus-Based Phrase Examples

The examples of phrases from both ROGER and ROGER_AI show (see Appendix 1) some overlap but also significant variation in how novice writers approach phrase construction in Romanian. In the case of adjectival modifiers (Amod), both datasets include complex, context-specific phrases such as *califat islamic* and *lider carismatic*. These phrases reflect the academic register and are typical of higher-level academic writing, which often includes descriptors for abstract political concepts. The inclusion of phrases like *știre fals* and *orient mijlociu* shows that students are dealing with current, real-world issues and political topics in their writing. In terms of adverbial modifiers (Advmod), the data highlights a clear difference between the two sets. In ROGER, phrases like *merge înapoi* and *aduce aminte* reflect more conversational or less formal structures, whereas ROGER_AI contains more sophisticated adverbial phrases such as *schimba radical* and *crește constant*, which reflect a more formal and academic tone. This indicates that the Romanian students' writing, which initially relies on simpler adverbial constructions, can become more complex through AI rewriting. The

direct object (Dobj) examples further support this trend. In ROGER, phrases like *atinge apogeu* and *atrage atenție* demonstrate a basic level of sophistication but remain relatively straightforward in construction. However, in ROGER_AI, phrases such as *prinde contur* and *întâmpina dificultate* are more nuanced and reflect a deeper engagement with abstract ideas, which is essential in academic discourse.

The phrases in EXPRES and EXPRES_AI corpora (see Appendix 2) demonstrate a higher level of phraseological sophistication compared to novice writers. In adjectival modifiers, phrases like *taxă vamal* and *competiție acerb* from EXPRES highlight precise, field-specific terminology used by experts. The AI-generated corpus, EXPRES_AI, includes similar phrases such as *inteligentă artificial* and *viziune holistic*, showing the AI's ability to replicate specialized language. For adverbial modifiers, expert writers in EXPRES use phrases like *menționa anterior* and *întâlni frecvent* to reference research methodology or past findings. The AI-generated EXPRES_AI corpus mirrors this with phrases like *gestiona eficient*, though with slightly less diversity. In direct objects, phrases such as *primi premiu* and *desfășura activitate* are commonly used in EXPRES, reflecting academic tasks. The EXPRES_AI corpus includes similar combinations like *rezolva problemă*, but occasionally simplifies the language. While the AI-generated texts replicate expert writing patterns, they generally display less variety and nuance compared to human-written texts.

ChatGPT-generated phrases in academic writing exhibit clarity, formality, and precision, but often lack the linguistic diversity seen in human-authored texts. For adjectival modifiers (Amod), phrases reflect a polished but predictable style. In adverbial modifiers (Advmod), straightforward combinations are frequent, prioritizing functionality over variety. For direct objects (Dobj), ChatGPT often uses structured verb-noun pairs showing a preference for familiar, formulaic expressions. While these phrases are contextually appropriate and academically sound, they tend to be more repetitive and less nuanced compared to human writing.

5. Discussion

The comparison between human writing and AI-rewritten texts based on the corpora reveals distinct differences in linguistic features. While the AI-rewritten texts (created by ChatGPT) exhibit improvements in some aspects of phraseological sophistication, they also demonstrate a noticeable reduction in phraseological diversity. Specifically, the AI tends to produce more polished and structured direct object phrases, which are typically more sophisticated in

comparison to the human-written texts. However, this comes at the expense of diversity in other syntactic structures, such as adjectival and adverbial modifiers.

In the human-written corpora (ROGER and EXPRES), there is greater variety in the construction of phrases, with more frequent use of less formulaic and more creative combinations. This diversity is particularly apparent in the use of adverbial modifiers, where human authors introduce more varied contextual elements. By contrast, the AI-generated texts tend to rely on a smaller set of syntactic patterns, resulting in more repetitive and predictable constructions. For example, while the AI produces accurate, context-appropriate phrases like *anticipa nevoie* or *rezolva problema*, it lacks the range of expressions seen in the human-written versions, where authors often employ a wider array of synonymous or alternative structures.

The AI-rewritten texts also show a tendency toward more compact, less verbose outputs. This often leads to the omission of subtler linguistic features, such as nuanced adverbial expressions or more complex syntactic dependencies that are frequently found in human writing. Additionally, while the AI can replicate the tone and formal structure of academic writing, its outputs occasionally lack the richness and depth characteristic of human authorship, particularly in areas that require more creative or interpretive language use.

In conclusion, the primary linguistic distinctions between human and AI-rewritten texts lie in the balance between phraseological sophistication and diversity. AI-rewritten texts, while capable of producing more refined and grammatically accurate phrases, tend to do so at the cost of syntactic variety and complexity. These findings suggest that AI rewriting models like ChatGPT are more efficient at producing structured, formulaic academic language but may underperform in generating the diverse linguistic features that are typically present in human-authored texts.

6. Conclusions

For rewriting essays (ROGER), ChatGPT is particularly effective at improving the clarity, structure, and formal tone of the text. It can polish simpler, more formulaic student writing by enhancing grammatical accuracy and producing well-formed, coherent sentences. In this context, ChatGPT excels at improving phraseological sophistication, such as using more appropriate adjectival modifiers (Amod) and refining direct object (Dobj) combinations. However, its limitations become evident when a more nuanced or creative expression is required, as it tends to simplify adverbial structures and reduce diversity, potentially making the rewritten essays more rigid and less varied in language.

For rewriting scientific articles (EXPRES), ChatGPT demonstrates some potential in handling the formal structure and specific terminology common in

academic papers. It can effectively maintain the precision of adjectival modifiers (e.g., *viziune holistic*) and ensure technical terms are used appropriately. However, in scientific writing, where depth of argumentation and subtlety in phrasing are crucial, ChatGPT's tendency to reduce adverbial modifier (Advmod) variety and repeat familiar structures may lead to less dynamic and engaging texts. These findings highlight both the strengths and limitations of AI in supporting different levels of academic writing in the Romanian language.

References

- Băniceru, C., and Tucan, D. (2018). Perceptions about “good writing” and “writing competences”: Romanian academic writing practices – A questionnaire study. In M. Chitez, C. I. Doroholschi, O. Kruse, L. Salski, and D. Tucan (Eds.), *University writing in Central and Eastern Europe: Tradition, transition, and innovation* (p. 103-112). Springer.
- Bercuci, L., and Chitez, M. (2023). The influence of the Romanian academic style on student writing in English. In M. Vranova (Ed.), *New to the LSP classroom? A selection of monographs on successful practices* (p. 169-185). Vernon Press.
- Bucur, A. M., Chitez, M., Mureșan, V., Dincă, A., and Rogobete, R. (2022). Ro-AWL: Building the first Romanian Academic Word List through corpus analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (p. 4739-4746).
- Chitez, M., Bercuci, L., Dincă, A., Rogobete, R., and Csűrös, K. (2021). Corpus of Romanian Academic Genres (ROGER). West University of Timisoara. Available at <https://roger-corpus.org/>
- Chitez, M., Rogobete, R., Mureșan, V., and Dincă, A. (2022). Corpus of Expert Writing in Romanian and English (EXPRES). West University of Timisoara. Available at <https://expres-corpus.org/>
- Chitez, M. 2022. *Corpus of Romanian Academic Genres (ROGER)* [Corpus]. Slovenian Language Resource Repository CLARIN.SI. <https://hdl.handle.net/11356/1700>
- Dincă, A., Bucur, A.-M., and Chitez, M. (2024). Tracking phraseological inter-complexity: A corpus-based contrastive analysis of learner English and Romanian novice academic writing. Paper presented at the *Learner Corpus Research Conference 2024*, Tartu, Estonia, September 5-7, 2024.
- Doroholschi, C. I., Băniceru, C., Bercuci, L., and Chitez, M. (2025, in press). Construction of voice in Bachelor theses: Between student writing and supervisor expectations. In G. Skar, J. Marine, and P. Rogers (Eds.), *Writing research across borders: From early literacy to learning to writing in professional life*. Colorado: WAC Clearinghouse.
- Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6(3), 241-252.
- Ene, E., and Mitrea, A. (2013). EFL writing teacher training beliefs and practices in Romania: A tale of adaptation. *The European Journal of Applied Linguistics and TEFL*, 2(2), 117-138.
- Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47, 583-621.
- Mititelu, V. B., Irimia, E., and Tufis, D. (2014). CoRoLa – The reference corpus of contemporary Romanian language. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 1235-1239.

- Mureșan, V., Rogobete, R., Bucur, A. M., Chitez, M., and Dincă, A. (2022). Phraseology in Romanian academic writing: Corpus-based explorations into field-specific multiword units. In D. Anca, M. Chitez, L. Dinu, and M. Dobre (Eds.), *Recent advances in digital humanities. Romance language applications* (p. 29-48). Peter Lang.
- Oravițan, A., Chitez, M., Bercuci, L., and Rogobete, R. 2022. Using the bilingual Corpus of Romanian Academic Genres (ROGER) platform to improve students' academic writing. In B. Arnbjörnsdóttir, B. Bédi, L. Bradley, K. Friðriksdóttir, H. Garðarsdóttir, S. Thouéšny, and M. James Whelpton (eds.), *Intelligent CALL, granular systems and learner data: Short papers from EUROCALL 2022*, pages 315-321. Research-publishing.net.
- Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, 15(1), 29-43.
- Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1), 121-145. <https://doi.org/10.1177/0267658317694221>
- Tufis, D., Mititelu, V. B., Bozianu, L., and Mihaila, C. (2006). Romanian WordNet: New developments and applications. *Proceedings of the 3rd Conference of the Global WordNet Association*, 337-344.
- Vandeweerd, N., Housen, A., and Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A partial replication study. *International Journal of Learner Corpus Research*, 7(2), 197-229. <https://doi.org/10.1075/ijlcr.20010.van>

Appendices

Appendix 1: Academic Phrase Examples per Linguistic Complexity Category in ROGER and ROGER_AI

Combinations with PMI>3 from ROGER

Amod	Advmod	Dobj
direcția general	aduce aminte	înclina balanță
panou solar	merge înapoi	cere scuză
orientului mijlociu	crește repede	atrage atenție
califat Islamic	lucra împreună	atinge apogeu
extremă stâng	crește considerabil	întâmpina dificultate
taxă vamal	influența pozitiv	atrage atenție
pol opune	datora mult	acorda atenție
armată roșii	discuta jos	crea premis
învățământ preuniversitar	deveni evident	juca rol
corporație transnational	interpreta corect	depune efort
scară planetar	răspunde rapid	pune capăt
lider carismatic	prezenta detaliat	capta atenție
închisore comunist	face posibil	lua prizonier
orientare sexual	face rău	desfășoara activitate
armată roșie	prezenta anterior	simți nevoie
război rece	întâmpla așa	purta denumire
atac terorist	exista încă	depăși așteptare
flux migrator	exprima liber	pune accent
avantaj competitiv	prezenta jos	pune amprentă
proprietate intelectual		suferi modificare

Combinations with PMI>3 from ROGER_AI

Amod	Advmod	Dobj
califat islamic	schimba radical	înclina balanță
inteligență artificial	întâlni des	prinde contur
știre fals	crește constant	atrage atenție

Amod	Advmod	Dobj
învățământ preuniversitar	întâlni frecvent	atinge apogeu
corporație transnațional	angaja personal	întâmpina dificultate
lider carismatic	apărea rar	juca rol
vârstă fraged	exprima liber	pune capăt
orientare sexual	utiliza frecvent	capta atenție
armată roșie	folosi frecvent	primi premiu
energie regenerabil	apărea adesea	desfășoara activitate
război rece	dezvoltat puțin	purta denumire
atac terorist		depăși așteptare
avantaj competitiv		rezolva problemă
proprietate intelectual		câștiga teren
regim totalitar		asuma responsabilitate
libertate cetățenesc		transmite mesaj
germania nazist		mobiliza resursă
învățământ superior		atinge obiectiv
regim autoritar		face apel
regim dictatorial		îndeplini criteriu

Appendix 2: Academic Phrase Examples per Linguistic Complexity Category in EXPRES and EXPRES_AI

Combinations with PMI>3 from EXPRES

Amod	Advmod	Dobj
taxă vamal	menționa anterior	primi premiu
competiție acerb	vedea jos	desfășoara activitate
trafic rutier	întâlni frecvent	suferi modificare
guvernanță corporativ	întâlni rar	acoperi gamă
transport aerian	diferi semnificativ	face referire
scară larg	prezenta anterior	aloca resursă
atlas lingvistic	afla încă	purta nume
gamă larg		presupune parcurgere
ordine alfabetic		atinge obiectiv
act normativ		desfășura activitate

Amod	Advmod	Dobj
teritoriu dacoromân		lăsa urmă
regat unit		reduce cost
viziune holistic		folosi algoritmi
stare emoțional		conține informație
modificare adus		vedea figură
parte integrantă		valorifica resursă
descriere detaliat		lua decizie
uniune european		produce efect
telefon inteligent		reglementa organizare
administrație public		ridica problemă

Combinations with PMI>3 from EXPRES_AI

Amod	Advmod	Dobj
inteligentă artificial	lucra împreună	juca rol
taxă vamal	menționa anterior	face abstracție
transport aerian	gestiona eficient	desfășoara activitate
atlas lingvistic	întâlni frecvent	rezolva problemă
act normativ	afecta negativ	face referire
teritoriu dacoromân	probabil puțin	purta nume
regat unit		atinge obiectiv
viziune holistic		anticipa nevoie
decizie luat		facilita comunicare
modificare adus		lăsa urmă
parte integrantă		reduce cost
descriere detaliat		folosi algoritmi
limbii român		indica fapt
zonă rural		respecta cerință
uniune european		vedea figură
intenție comunicativ		investigha mod
proces decizional		lua decizie
academie român		îndeplini obiectiv
format electronic		conține informație
limbă romanic		permite accesare

CHAPTER 3

APPLICATIONS

REDDIT IS ALL YOU NEED: AUTHORSHIP PROFILING FOR ROMANIAN

ECATERINA ȘTEFĂNESCU, ALEXANDRU-IULIUS JERPELEA

“Tudor Vianu” National College of Computer Science, Bucharest, Romania
ecaterina@stefanescu.eu, alex.jerpelea@gmail.com

Abstract

Authorship profiling is the process of identifying an author’s characteristics based on their writings. This centuries old problem has become more intriguing especially with recent developments in Natural Language Processing (NLP). In this paper, we introduce a corpus of short texts in the Romanian language, annotated with certain author characteristic keywords; to our knowledge, the first of its kind. In order to do this, we exploit a social media platform called Reddit. We leverage its thematic community-based structure (subreddits structure), which offers information about the author's background. We infer a user's demographic and some broad personal traits, such as age category, employment status, interests, and social orientation based on the subreddit and other cues. We thus obtain a 23k+ samples corpus, extracted from 100+ Romanian subreddits. We analyse our dataset, and finally, we fine-tune and evaluate Large Language Models (LLMs) to prove baselines capabilities for authorship profiling using the corpus, indicating the need for further research in the field. We publicly release all our resources.¹

1. Introduction

Author-specific traits such as demographic characteristics (age, gender, educational background, etc.) (Akçay, 2017), personality traits (Ahmed and Naqvi, 2015), interests, and disciplinary background (Coffin and Hewings, 2003) all influence the way humans communicate and reflect in the language used. This includes written text, which this paper aims to study.

Authorship profiling is one of the three categories within the broader field of authorship analysis (the other two being authorship attribution and verification) (Misini et al., 2022) with the purpose of building the profile of a text’s author (Lundeqvist and Svensson, 2017). This can be done by analysing the information extracted from a text through the following approaches (Pervaz et al., 2015; Reddy et al., 2016): (1) stylometric (Grivas et al., 2015), which analyses writing style,

¹ <https://huggingface.co/roship-profiling>, Accessed: 2024-10-12

(2) content-based, (3) topic-based, and (4) mixed. The characteristics that one may want to identify about the author can vary, depending on the nature of the task. They can range from demographic characteristics (Reddy et al., 2016), such as age, gender, and native language, to psychographic characteristics (Pervaz et al., 2015), such as extraversion or neuroticism. We deploy a mixed approach that aims to identify as many keywords from the described categories (if possible), relying on Large Language Models (LLMs) to extract them. We aim that the LLM learns this behaviour from the very nature of our collected dataset.

Social media has become an essential medium of modern communication, enabling people to share opinions and content within specific contexts (Schrape, 2016). In today's world, it has become a vital platform for human interaction, making it easier for people with common interests and traits to connect. Among these spaces, Reddit is one of the best at representing these aspects, as users can share interests and engage with like-minded individuals (Green and McCann, 2022). This is possible due to the platform's structure, functioning on the basis of subreddits, self-formed communities of users brought together by a certain topic, interest, and, most importantly, certain character traits (Medvedev et al., 2019; Green and McCann, 2022). By manually constructing a set of general characteristics that define a certain subreddit's users, we are able to annotate posts with certain keywords that relate to that user, otherwise named as the author.

Through this paper, we aim to advance the capabilities of authorship profiling technology for the Romanian language. To our knowledge, no such systems have been specifically designed for the Romanian language.

More than that, we note the scarcity of valuable data in this regard for the Romanian language. We have addressed this issue by building our own dataset, based on Reddit posts within subreddits, using their corresponding topics to infer useful information about the authors. We have decided to analyse the text from the perspective of the following categories: (1) subdialect of the Daco-Romanian dialect, (2) educational or employment status, (3) personal interests, such as computer science and music, and (4) personal inclinations, such as social orientation and technological interest.

We then finetune a LLM for the task. However, given that the attribution of a clear attribute for each category is not always possible, the LLM can be rather viewed as an author-related keyword generator.

Our work to fill the existing gap can thus be summarised as follows:

A corpus of Reddit posts, comments, and other texts in the Romanian language, annotated with keywords, that correspond to certain categories, and which relate to their author's background/characteristics.

A fine-tuned Llama 3.1-8B Instruct model (Dubey et al., 2024), which generates such keywords from text.

In this paper, we first describe the data collection/annotation process and our baseline LLM development. In the end, we discuss the limitations of our low-resource background and the need for further research in the emerging field of Romanian authorship profiling.

2. Related Work

With recent technological developments, authorship profiling has evolved in the past years (Chinea-Rios et al., 2022), reaching and developing for languages different from English. There are various authorship profiling applications for a diverse set of languages, including Arabic (Bassem et al., 2024), Spanish (Villegas et al., 2014; García-Díaz et al., 2024), Lithuanian (Kapočiūtė-Dzikienė et al., 2015), and Russian (Litvinova et al., 2016.). In most cases, the tasks involve the determination of age and gender (HaCohen-Kerner, 2022) or other demographic traits (Reddy et al., 2016). Usually, if the systems target personality traits (Ahmed and Naqvi, 2015), they are usually regarding the five mainly researched ones, namely extroverted, stable, agreeable, conscientious, and open (Iqbal et al., 2015). Research regarding personal inclinations is not as common. A serious issue impacting all approaches is the lack of available data on the correlation between an author’s characteristics and his written text. This causes even more concern for less-resourced languages.

To our knowledge, there are not any approaches that leverage Reddit for an authorship profiling related corpus construction. One of the most notable authorship analysis attempts focuses on the use of other social media platforms related to authorship attribution (Casimiro et al., 2020; Matias and Digiampietri, 2023).

A significant work involving Romanian and Reddit is the first dataset targeting the popularity prediction of Romanian Reddit posts (Rogoz et al., 2024). Our corpora’s building steps share similarities, but our approaches differ. For example, when scrapping Reddit, we only used Reddit’s API. This API has a limit of 1000 posts per subreddit. Rogoz et al. (2024) have overcome this limitation by using an open-source archive. The numbers of data samples of the dataset and our corpus are very close (28107 and 26517, respectively). However, the ratio between the number of subreddits considered and the number of posts extracted per subreddit is very different. Rogoz et al. 's (2024) corpus only targets 5 Romanian subreddits, while we tackle 146.

Another noteworthy development in authorship analysis involving the Romanian language involves authorship profiling’s more popular counterpart, authorship attribution (Nitu and Dascalu, 2024).

Lastly, Ciobanu and Dinu (2016) have tackled automatic dialect identification for Romanian. Our LLM, besides other author characteristics, does a similar task. However, we only identify varieties of Daco-Romanian and do not approach the broader Romanian dialectal continuum.

3. Methodology

3.1. Finding Romanian Subreddits

Due to the nature of tag attribution to Reddit posts that belong to a certain subreddit, establishing a list of valid subreddits is an important step. After the scouting process, we have identified a list of 263 Romanian subreddits. Among these, 159 were manually distinguished through general searches or sidebar suggestions, while the other 104 were obtained with the help of a Reddit exploration tool that identifies a network of subreddits based on an input subreddit name. A chosen subreddit must be relevant and centred toward the certain topic(s) or traits that most of its authors have, in order to make the tagging process effective. Other criteria were the number and the quality of the posts. Candidates not meeting these requirements were removed. This process results in a total of 146 subreddits.

3.2. Characterising Subreddits

3.2.1. Subdialects

This tag was given to subreddits corresponding to Romanian cities and regions, where upon a manual analysis, speakers are likely to write in a region-specific register. We have taken into consideration four big varieties of Daco-Romanian: Transylvanian, Moldavian, Oltenian, and Wallachian. As can be seen in Table 2, the distribution is somewhat uniform, except for the Oltenian subdialect which has significantly fewer posts. In Table 1 are the subreddit names that have been assigned to each subdialect (they mostly correspond to cities and regions).

Table 1: Attribution of regional speech variety to various subreddits

Daco-Romanian variety	Subreddits
Transylvanian subdialect	albaiulia, Arad, bistrita, cluj, Jobs_Timisoara, mures, Oradea, satumare, TarguMures, timisoara, Timisoara_night_life, Zalau
Moldavian subdialect	Bacau, Botosani, iasi, moldova, Moldoveneste, Moldovenii, PiatraNeamt, Suceava

Daco-Romanian variety	Subreddits
Wallachian subdialect	bragadiru, brasov, bucuresti, constanta, pitesti, Ploiesti
Oltenian subdialect	Craiova, Oltenia

Table 2: The distribution of data samples targeted by the subdialect tag within its categories

Subdialect	Number of data samples
Moldavian	1744
Transilvanian	2632
Wallachian	2200
Oltenian	361

Table 3: The distribution of data samples targeted by the status tag within its categories

Status	Number of data samples
Pre-University students	965
University students	2670
Employed	2247

3.2.2. Status

This type of tag refers to the author’s education or employment status, and because they account for different stages in life, they can be related to some broad age brackets. The characteristics of the subreddits assigned with this type of tag were, for each case, the following:

- Pre-university students – subreddits corresponding to a high school or to the Romanian baccalaureate, the exam to graduate high school,
- University students – subreddits corresponding to certain academic fields, universities, and educational apps,
- Employed – subreddits corresponding to jobs and employment.

In this case, the information regarding the first category is not as prevalent as the other two, due to the nature of the subreddits taken into consideration for analysing, as Table 3 demonstrates.

3.2.3. Labels and Field of Interest

As subreddits are commonly centred around a certain concept, the users within that community share an affinity for the topic. Based on almost every subreddit’s context, we have identified one or more labels that indicate a certain general interest displayed by the users (authors). We obtain a number of 32 labels that are assigned to each post based on the subreddit that it originates from. Together, these

topics form a very diverse set of subjects, but they can also be closely related, like mathematics and physics, or conservatism and liberalism. For such cases, besides the interests of the authors, we have come up with a new category: a *field of interest*. Only 6 in number, their purpose is to group similar labels into broader classes that once again indicate a more general interest of the author. Table 4 showcases this grouping. In case the attribution of specific labels fails, then we aim to eliminate mischaracterization at least with the *field of interest* broader approach.

Table 4: An illustration showcasing the relations between labels and fields of interest

ART design fashion literature music	SCIENCE chemistry physics informatics mathematics medicine psychology	TECHNOLOGY cars motorcycles	POLITICS conservatism liberalism	ECONOMY banking crypto work	SPORT exploration football fishing
ADDITIONAL: animals, dating, environment, movies, philosophy, gardening, history, games, food, parenting, partying, chess					

3.2.4. Personal Inclination

Moreover, with the help of the *field of interest* category, we can construct *personal inclination* tags. These tags aim to encompass every aforementioned label and are adapted to be more suitable for assigning attributes to an author's profile. Multiple such tags can be assigned to a post, based on how its corresponding subreddit has been characterised. This results in 12 types: active lifestyle, arts-oriented, hobbyist, interested in economy, interested in games, interested in health, interested in history or politics, interested in nature, interested in science, interested in technology, media enthusiast, and socially oriented.

We have to acknowledge these tags are more related to the text's content, but their correlation to the author's background is important.

3.3. Reddit Post Extraction

The posts were gathered using the Reddit API through the Python Reddit API Wrapper (PRAW) library. This method implies a limitation of up to 1000 posts per subreddit or per subreddit category. We aim to keep a balanced dataset anyway so this is not an obstacle. The script checks each post for text content. The information we extracted for each post consists of the title and the body of the post (if it exists). For posts without a body, the title is all that remains. It is added to the dataset only if it has at least 75 characters. We settled on this restraint to ensure that it conveyed a message and that it was related to a certain topic. Once

all the relevant information is collected, it is compiled into a CSV file. The resulting dataset has 29,317 annotated posts.

3.4. Dataset Cleaning

Posts in the English language are one of the main problems we encountered with the first version of the dataset, compromising our approach aimed at the Romanian language. To eliminate them, we used Google’s `langdetect` library to separate between different languages through two approaches: (1) identifying Romanian text (25512 rows) and removing the rest (1805 rows), and (2) identifying English text (1256 rows) and eliminating them, resulting in a 28060 rows dataset. Reddit posts are informal in nature, as Romanian written text can oftentimes be mixed with English words or phrases. We chose to be tolerant in this case, and upon manual inspection, the second method yielded better results for our purpose.

Additionally, we have removed excessive whitespaces and compressed multiple consecutive newlines to a maximum of two. Then, we removed URLs, by identifying patterns generally associated with web addresses. If the text of the body consisted only of URLs, we checked the length of the title once again and kept the row in the dataset only if it had a minimum of 75 characters. Stray symbols and rogue parentheses were eliminated as well. After these modifications, the dataset consists of 26,517 data samples. In order to preserve as much linguistic information as possible, we chose to not make any more changes to the dataset.

3.5. Final Dataset Construction

After crawling the Reddit posts, note that sometimes, we are not able to generate keywords for each of the presented categories (*subdialect*, *status*, *interest labels*, *field of interest*, *personal inclination*). For example, it is impossible to infer for certain subreddits what regional speech is used there.

Seeking to mitigate this issue and provide a more simplistic structure for the LLM to learn, we merge all keywords from each category into a single set of author-related keywords. The keywords do appear in the same order described in this paper. With this less structured approach, we also shift our authorship profiling task from a classification problem to a feature extraction one.

4. Authorship Profiling

Followingly, we use the processed corpus to build a baseline authorship profiling tool that can gather key aspects about a text’s author in the form of keywords. As mentioned in Section 3.5, these keywords are unorganised, and are generated only when relevant. For example, if there cannot be said anything about the subdialect a writer uses, such keywords should not be generated.

4.1. Model Training

We leverage Large Language Models (LLMs), which have proven useful in numerous NLP tasks (Kocmi et al., 2024). Specifically, we seek to utilise their capabilities for feature extraction tasks. The LLM we train is given the task of extracting author-specific keywords.

We experimented with the Llama 3.1-8B-Instruct (Dubey et al., 2024). It is the most recent LLaMA model, and it covers a diverse set of instructions. Romanian language, although not officially supported, is covered in about 15% of the multilingual tokens from the dataset used in pre-training.

Given its zero or few-shot learning capabilities, we expect the model to perform well even in the conditions of parameter-efficient finetuning. Specifically, we employ LoRA (Low-Rank Adaptation of Large Language Models) fine-tuning (Hu et al., 2021). Instead of updating all model parameters, LoRA focuses on fine-tuning only a small set of parameters through the decomposition of low-rank matrices, which are later added to the model weight matrices.

We are thus able to afford computation, by training around 10% of the number of parameters which would usually be updated during a full fine-tune of the Llama 3.1-8B-Instruct. This has both benefits, as the model will diverge less from its base version (will not “forget” important pre-trained knowledge), and disadvantages, as it is able to adapt less to the given task (Biderman et al., 2024).

We split out the dataset into three pieces: *train*, *eval*, and *test* with a 95:5:5 ratio respectively. The first is used for training our model, the second is used for in-training loss evaluation, and the latter is used in the test section, where we showcase the final loss of our model.

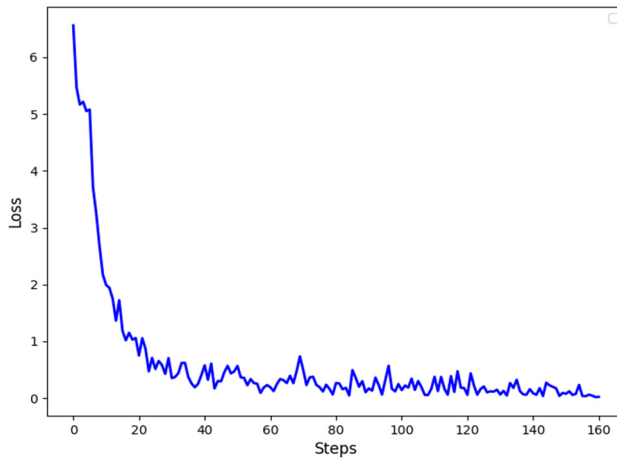


Figure 1: Loss over time while training, measured on the evaluation split of the dataset

We showcase the prompt format of the LLaMA model in Table 5. The model is trained to complete the text in continuation of the “assistant” tag. Hyper-parameters are also available in Table 5. Note that we stop training after 1 epoch as loss from measured on the evaluation split converges rather quickly, and ultimately plateaus (Figure 1). Also note that we dispense samples with a length of over 2200 tokens (measured with the model’s BPE tokenizer), because of the limited GPU memory we have available.

Table 5: Prompt format example for Llama 3.1-8B-Instruct (actual text and keywords may vary), and hyperparameters used for training

System	< start_header_id >system< end_header_id > Ești un asistent util care analizează texte și autorii lor.< eot_id >
User	< start_header_id >user< end_header_id > >Generează cuvinte-cheie din următorul text, care să aibă legătură atât cu conținutul textului, cât și cu trăsături ale autorului. Puteți, va rog, sa mi explicați teorema cosinusului pe scurt, sunt în clasa a 7a și vreau sa învăț cât mai repede și efficient < eot_id >
Model	< start_header_id >assistant< end_header_id > fizica, stiinta, interes stiintific

Hyperparameter value	Value
Effective Batch Size	8
Number of epochs	1
Learning rate	0.0002
Warmup steps	10
Weight decay	0.0
Maximum sequence length	2200
LoRA rank	16
LoRA alpha	16
LoRA dropout	0.05

4.2. Model Testing

On the *test* split, we employ two ways of measuring our model’s success: exact match and Intersection over Union (IoU).

First, for every evaluated sample, we compare the set of generated keywords by our model, and the set of expected keywords (according to the corpus). We obtain an accuracy of 78.5%.

Second, we acknowledge that our corpus annotation might be prone to errors, in the sense that some keywords may not be evidently related to the text just based on the subreddit section, and thus are very hard for our model to infer. Consequently, we apply a more forgiving metric, measuring the ratio between the cardinality of the intersection and the union of the two sets of keywords (predicted and expected). On average, on the *test* split, we obtain an 80.4% IoU average.

4.3. Other Model Statistics and Behaviours

We observe the model’s tendency to generate slightly fewer keywords per sample. The reference samples yield a 2.31 average of keywords, whilst the generated outputs yield a 2.26 average.

Moreover, there are 53 unique expected keywords in the *test* split. Putting together all of our model’s predictions, we obtain 52 unique generated keywords; all of them are included in the expected corpus keywords, proving that the model is not yet able to generate new types of information (that would have not been encountered in the fine-tuning phase, but would have rather been inferred from its pre-training phase) about an author.

Based on this paper’s authors’ manual review, we make the observation that the model tends to generate slightly more content-focused keywords, rather than authored-focus. This pattern appears in our dataset as well but to a lesser degree. We are unable to quantify this phenomenon, but we note that the difference is not major, as authorship profiling remains the model’s main focus.

Lastly, we observe the model’s inability to combine various scenarios from the training phase. For example, in our dataset, most of the samples which flag the author as a speaker of a certain subdialect, do not have any other additional keywords included, due to the limitations of our annotation methodology. This behaviour is also observed in the model’s generated keywords. Predictions related to the spoken subdialect are rarely accompanied by other keywords. For instance, the model generates “grai ardelenesc” (EN: *Transylvanian subdialect*) about the author of the following text: “Gulas bun in Budapesta. Intreb clujenii ca mno e o comunitate mai mare de cunoscatori” (EN: *Good stew in Budapest. I ask the people of Cluj because there is a larger community of connoisseurs*). However, the keyword “food interest”, which is found in some other samples of our dataset, could have been suitable to be generated, but it is not.

All in all, the accuracy and IoU are promising for this baseline approach, and suggest how further research about Romanian authorship profiling has a lot of potential if further research is done.

5. Conclusions and Future Work

This paper introduced the first authorship profiling-oriented corpus for the Romanian language, with text samples annotated with keywords from various categories. Moreover, we train a baseline LLM for a keyword generation task, a problem that can also be viewed as authorship feature extraction. The promising, yet limited, capabilities of the model suggest how more research is needed in this field to obtain better results. In the future, we aim to obtain a more diverse dataset, that is not only sourced from Reddit. More robust methodologies for data annotation should be employed, as our approach is prone to some errors (limitations). Lastly, other model architectures, sizes, and fine-tuning methodologies should be experimented with.

6. Limitations and Ethical Concerns

First, the tags were manually extracted based on the characteristics of the subreddit. This is a highly biased process, susceptible to personal assumptions and stereotypes. Although not documented, this could potentially generate unwanted toxic behaviours in our model. Moreover, the process of annotating almost all subreddit posts with the same tags is prone to error. Sometimes, certain tags (specific to a given subreddit) are impossible to infer by a model from the given text. More thoroughness, diversity, and manual evaluation are needed to mitigate these issues.

Secondly, although our classification is diverse (tackling demographic traits and personality characteristics at the same time), it comes with its downsides. Keywords much more closely related to the characteristics of the author, such as the status, get intertwined with keywords targeting just the text's content. Oftentimes, the line between the two can get blurry and hard to distinguish.

Lastly, our computing resources are limited, so we were not able to perform extensive experiments with other models, or even full fine-tuning.

References

- Ahmed, Javeria, and Irum Naqvi. 2015. Personality traits and communication styles among university students. *Pakistan Journal of Social and Clinical Psychology*, 13.2, p. 53-59.
- Akçay, Ahmet. 2017. Examination of the Relationship between Demographic Characteristics of the Family and the Language Development of Children. *International Journal of Higher Education* 6.5, p. 168-180.
- Bassem, Bsir, Ajith Abraham, and Zrigui Mounir. 2024. Transformer Models for Authorship Profiling in Arabic Social Media Texts. *International Journal of Computer Information Systems and Industrial Management Applications*, 16.3, p.13.
- Biderman, Dan, et al. 2024. Lora learns less and forgets less. arXiv preprint arXiv:2405.09673.

- Casimiro, Guilherme Ramos, and Luciano Antonio Digiampietri. 2020. Authorship Attribution using data from Reddit forum. *Proceedings of the XVI Brazilian Symposium on Information Systems*.
- Chinea-Rios, Mara, et al. 2022. Zero and few-shot learning for author profiling. *International Conference on Applications of Natural Language to Information Systems*. Cham: Springer International Publishing.
- Ciobanu, Alina Maria, and Liviu P. Dinu. 2016. A computational perspective on the Romanian dialects. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*.
- Coffin, Caroline, and Ann Hewings. 2003. Writing for different disciplines. In: Coffin, Caroline; Curry, Mary Jane; Goodman, Sharon; Hewings, Ann; Lillis, Theresa and Swann, Joan eds. *Teaching Academic Writing: A Toolkit for Higher Education*. London, UK: Routledge, p. 45-72.
- Dubey, Abhimanyu, et al. 2024. The llama 3 herd of models. arXiv preprint arXiv:2407.21783.
- García-Díaz, José Antonio, Ghassan Beydoun, and Rafel Valencia-García. 2024. Evaluating Transformers and Linguistic Features integration for Author Profiling tasks in Spanish. *Data and Knowledge Engineering* 151: 102307.
- Green, Daryl, and Jack McCann. 2022. The Rise of Reddit: A Case Study on Persuasive Technology. *American Research Journal of Business and Management*, vol. 8, no. 1, p. 24-28.
- Grivas, Andreas, Anastasia Krithara, and George Giannakopoulos. 2015. Author profiling using stylometric and structural feature groupings. *Proceedings of CLEF 2015: Conference and Labs of the Evaluation Forum Experimental IR meets Multilinguality, Multimodality and Interaction*. CEUR-WS.
- HaCohen-Kerner, Yaakov. 2022. Survey on profiling age and gender of text authors. In *Expert Systems with Applications* 199: 117140.
- Hu, Edward J., et al. 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
- Ifrah Pervaz, Iqra Ameer, Abdul Sittar, and Rao Muhammad Adeel Nawab. 2015. Identification of author personality traits using stylistic features. *Proceedings of CLEF*.
- Iqbal, Hafiz Rizwan, Muhammad Adnan Ashraf, and Rao Muhammad Adeel Nawab. 2015. Predicting an Author's Demographics from Text using Topic Modeling Approach. *CLEF Working Notes*.
- Kapočiūtė-Dzikienė, Jurgita, Andrius Utkā, and Ligita Šarkutė. 2015. Authorship attribution and author profiling of Lithuanian literary texts. *Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing*.
- Kocmi, Tom, et al. 2024. Preliminary WMT24 Ranking of General MT Systems and LLMs. arXiv preprint arXiv:2407.19884.
- Litvinova, Tatiana, et al. 2016. Ruspersnality: A Russian corpus for authorship profiling and deception detection. *International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*. IEEE.
- Lundeqvist, Elias, and Maria Svensson. 2017. Author profiling: A machine learning approach towards detecting gender, age and native language of users in social media. PhD thesis, Uppsala Universitet.

- Matias, Vinicius Alves, and Luciano Antonio Digiampietri. 2023. Authorship attribution of comments in Portuguese extracted from Reddit. *Revista Brasileira de Computação Aplicada*, 15.2: p. 1-10.
- Misini, Arta, Arbana Kadriu, and Ercan Canhasi. 2022. A survey on authorship analysis tasks and techniques. *SEEU Review* 17.2: p. 153-167.
- Nitu, Melania, and Mihai Dascalu. 2024. Authorship Attribution in Less-Resourced Languages: A Hybrid Transformer Approach for Romanian. *Applied Sciences* 14.7: 2700.
- Reddy, T. Raghunadha, B. Vishnu Vardhan, and P. Vijayapal Reddy. 2016. A survey on authorship profiling techniques. *International Journal of Applied Engineering Research* 11.5: 3092-3102.
- Rogoz, Ana-Cristina, Maria Ilinca Nechita, and Radu Tudor Ionescu. 2024. PoPreRo: A New Dataset for Popularity Prediction of Romanian Reddit Posts. arXiv preprint arXiv:2407.04541.
- Schrape, Jan-Felix. 2016. Social Media, Mass Media and the 'Public Sphere': Differentiation, Complementarity and Co-Existence. *SOI Discussion Paper*, No. 2016-01, Universität Stuttgart, Institut für Sozialwissenschaften, Abteilung für Organisations und Innovationssoziologie.
- Villegas, María Paula, et al. 2014. A Spanish text corpus for the author profiling task. *XX Congreso Argentino de Ciencias de la Computación*, Buenos Aires.

A MULTIDIMENSIONAL APPROACH FOR TEXT ANALYSIS

VLAD BELCIUG, ELENA PELICAN

*Ovidius University, Faculty of Mathematics and Informatics, Blvd. Mamaia 124,
Constanța, Romania*

vlad.belciug@365.univ-ovidius.ro
epelican@univ-ovidius.ro

Abstract

This paper introduces a novel system for text analysis based on a multidimensional design that provides a comprehensive and interpretable understanding of text characteristics. Unlike traditional methods, which often rely on surface-level or singular aspects to interpret text, our system evaluates texts across a diverse set of empirically grounded dimensions. These dimensions include Static, Dynamic, Affective, Cognitive, Time, Place, Animate, Inanimate, Quantitative, Qualitative, Generic, Specific, Social, Individual, Positive, Negative, Short-term, Long-term, Formality, Informality, Politeness, Directness, Novelty, Commonality, Intentionality, and Objectivity. Each dimension captures a unique aspect of language, allowing for a nuanced and structured analysis of complex texts, with the system allowing for additional dimensions to be added in the future. In addition to individual dimension scoring, the system incorporates the calculation of synergies between these dimension scores to reveal features that are not apparent when dimensions are examined in isolation. This capability is particularly valuable for applications such as discourse analysis, predictive modelling, and content evaluation. Designed to run locally, the system can operate in two modes: a high-performance mode leveraging free large language models (LLMs) for detailed analysis and a traditional mode utilizing rule-based methods and lightweight machine learning models for resource-constrained environments. Our approach provides interpretable, quantifiable metrics and clear visualizations that provide an “at-a-glance” understanding of the text. Furthermore, its local implementation allows for customization and ensures data privacy, making it suitable for a wide range of applications, from academic research to sensitive corporate analysis. By advancing the capabilities of text analysis tools and combining different approaches to analyse text, this system represents a significant step forward in understanding the multifaceted nature of language.

Keywords: action prediction, text analysis, multidimensional, natural language processing, synergies

1. Introduction

1.1. Background and Motivation

Natural language processing (NLP) is a subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language. NLP enables computers and digital devices to recognize, understand and generate text and speech by combining computational linguistics – the rule-based modelling of human language – together with statistical modelling, machine learning and deep learning (Cole, Jim 2024). Most traditional text analysis methods focus on a single dimension (e.g., sentiment polarity or syntactic structure) and therefore may miss the intricate interplay of many linguistic features that together carry meaning in a complex manner.

1.2. Originality and Contributions

Our system introduces a customizable multidimensional approach to text analysis, evaluating texts across multiple empirically grounded dimensions and synergies between them. The main idea is to quickly obtain an overall view of the text that allows the user to grasp at a glance the general nature of the text, not what the text is about exactly, but rather a complex but easy to understand description of it. Because of this goal, the detailed word-level analysis becomes less important, so that valency shifters or ambiguous words which could fall into different categories depending on the contextual sense, lose importance of impact over the final result. The inclusion of one word in more than one dimension is not necessarily a bad thing, since it helps paint the overall portrait of the text, meaning that the text has many aspects, is complex. This is why calculating synergies between dimensions can reveal deeper insights that are not apparent when dimensions are considered in isolation. By using the free versions of LLMs for advanced analysis, our proposed system tries to bring the power of this technology within the reach of the small and medium businesses and individual users, on a local environment. Since the direct user of the system does not necessarily have to be a human, the system can help automated programs initially sort texts according to general attributes before performing other tasks involving them. For humans, this system could be used in many different domains, from email sorting, to customer relations or advertising, to real-time chat analysis while using the resources of the offline environment, without having to give away the company data to online mediums. Moreover, the system is designed to operate in both high-performance and resource-constrained environments, adding versatility and ensuring users can benefit from detailed analysis regardless of the computational resources available.

2. Related work

This being a complex, interdisciplinary project, we have tried to reference more acknowledged works from the different domains related to our dimensional approach for text analysis, although the research among more recent 2023-2024 works is an ongoing process. Every domain of activity that relates to our system is vast (computational linguistics, psychology, artificial intelligence), thus making the identification of academic works that specifically refer to aspects of our approach a slow endeavour.

2.1. Multidimensional Text Analysis

Multidimensional text analysis has been explored in several contexts. For example, to name only a few, the multivariate analysis of linguistic variation by Biber (1988) is a ground-breaking study, that considers many characteristics of written and spoken differences. Also, the Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al. 2015) enables us to perform psychological text analysis along different types. Coh-Metrix, a computational tool designed to analyse texts at multiple levels, focuses on various linguistic and cognitive dimensions (Graesser et al., 2011).

2.2. Advances in Large Language Models

Advances in large language models (LLM), such as BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) have greatly enhanced the overall ability of NLP. Such deep learning models can develop complex linguistic patterns and generate text that is very similar to human produced texts, but tend to be black boxes, meaning that they offer little in the way of interpretability. Also, more complex LLMs are more computationally expensive making them very hard to use locally and so, they are usually used online, because they are hosted by their respective developing companies. As a result, the user must share his/ her data in order to use them, not to mention the fact that there are few to none ways to use more than one at a time and they do not offer a bird's eye view of the text corpus and visual representations, unless specifically prompted so and with limits regarding the size of the text data.

2.3. Our Approach

We believe that a system capable of integrating different machine learning techniques and LLMs under the same hood would give users a way to activate the best part of all, at the same time. Moreover, we believe that the user should be able to pick the tool he/she wants to use, this is why, going further, we plan to develop a more complex user interface that would allow him/her to further

customize the product. Our approach provides interpretable dimension-wise and quantifiable metrics over the dimensions used and provides a customizable tool that can be used locally and perfected by each user in time, thus tailoring it specifically for the purposes she wants to use it. This strategy offers both deep analysis and text explainability, neglecting the challenges of prior methods by combining the benefits of rule-based, machine learning and LLM approaches. Our system does not propose new NLP algorithms for each dimension specifically, but rather integrates existing free tools and libraries into a modular and customizable configuration.

3. Theoretical Foundations

3.1. Selection of Dimensions

Our choice of dimensions represents an example of how the system can be used, what we believe would be useful to find out about any text before delving deeper into it. We have looked into different disciplines when deciding what dimensions we could confidently describe for this functional prototype. Because each dimension should focus on a distinct aspect of language that contributes to the meaning and effect of a text, we have chosen linguistics, psychology, and communication to show us the way for selecting them. Regarding linguistics, Biber's work (1988) emphasizes the importance of analysing multiple linguistic features to understand language variation, psychology helped us align the dimensions with psychological constructs, such as emotion (affective), cognition, and social behaviour (Pennebaker et al., 2015), while politeness theory (Brown et al., 2020) and theories of formality (Heylighen and Dewaele, 1999) from communication studies enrich the dimensions related to interpersonal communication. We have developed these dimensions trying to capture the most important general aspects of text that would have relevance in an overall assessment. This being a customizable system, new dimensions may prove worth developing or some of the proposed ones may prove not as important as anticipated.

Our proposed dimensions are the following. *Affective*: to measure the emotional content expressed in the text. *Cognitive*: to assess the presence of cognitive processes such as thinking and reasoning. *Time*: to identify temporal references. *Place*: to detect spatial references. *Static*: to describe states or conditions that are not changing. *Dynamic*: to describe actions, processes, or changes. *Animate*: to reference living entities. *Inanimate*: to reference non-living objects. *Quantitative*: to show numerical data and measurements. *Qualitative*: to show descriptive, non-numerical language. *Generic*: to express the use of general, non-specific language. *Specific*: to express the use of detailed and precise language. *Social*: to

focus on group dynamics and collective experiences. *Individual*: to focus on personal experiences and individual perspectives. *Positive*: to reveal the presence of positive sentiment and language. *Negative*: to reveal the presence of negative sentiment and language. *Short-term*: to express the references to immediate or near-future timeframes. *Long-term*: to reference distant future or enduring states. *Formality*: to evaluate the use of formal language and structures. *Informality*: to express the use of informal, colloquial language. *Politeness*: to show the use of polite expressions and strategies. *Directness*: to evaluate the clarity and straightforwardness of communication. *Novelty*: to see if new ideas and concepts are introduced. *Commonality*: to assess the reference to familiar or widely known concepts. *Intentionality*: to evaluate the expression of intent or purpose. *Objectivity*: to measure the degree of impartiality and factual reporting.

3.2. Synergies Between Dimensions

Since each dimension expresses an aspect of the text, we have realized that, besides the fact that the entire ‘picture’ depicted by all of them has value as a whole, the distribution of high and low scores for some of the dimensions can hold meaning and complete the analysis. Synergies are combinations of dimension scores that, when occurring together, could reveal deeper insights into the text. The theoretical basis for synergies comes from the understanding that language features often interact to produce compound effects on interpretation and reception. The synergies we have envisioned between various dimensions are identified and explained based on the scores of each dimension exceeding a certain threshold. The synergies we present are only examples of possible insights that could appear as relevant for the users going forward.

4. System Architecture and Methodology

Our system is designed to be modular and extensible, capable of analysing text across multiple dimensions and identifying synergies. Because it is designed to run locally, with no internet connection, local resources must be taken into consideration. This is why we have designed the system to first assess the local resources (processor, memory, etc.) and inform the user regarding which of the two methods of analysing text is more suitable for his/her hardware configuration (see Figure 1). The two methods are: a *high-performance mode*, which leverages free LLMs for detailed analysis, suitable for environments with ample computational resources and a *standard mode*, which uses rule-based methods, word lists and lightweight machine learning models, ideal for resource constrained environments. The system is designed to split the target file/s into chunks if the user wants to proceed even if the resources are low. The user is free to use

whichever method he/she wants, knowing that choosing the advanced method while having low resources will take a lot more time.

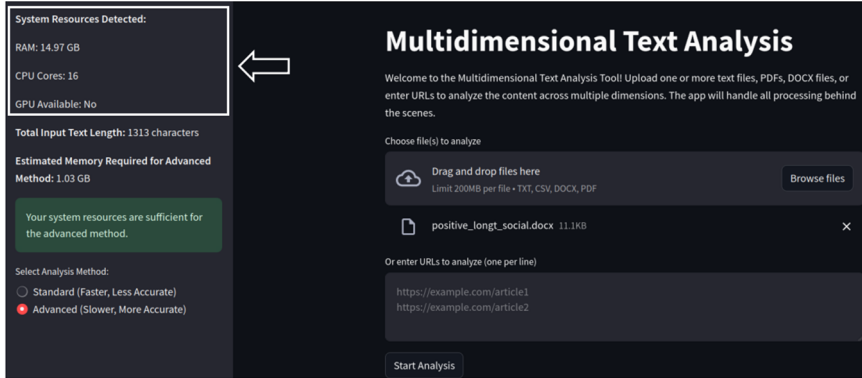


Figure 1: The resources area

The system architecture allows for easy integration of new dimensions and methods, ensuring adaptability to the evolving research needs of the user. According to the option of the user, dimensions can be calculated using either standard or advance methods.

The **standard methods** rely on traditional computationally efficient machine learning techniques and linguistic approaches. These methods are used to analyse dimensions by employing simpler models like NLTK, TF-IDF, Bag of Words (BoW), rule-based approaches or word lists.

The **advanced methods** incorporate free versions of large language models (LLMs), such as those provided by Hugging Face's GPT-2, GPT-Neo, or DistilBERT, to provide more sophisticated and in-depth analysis. These models leverage deep learning and transfer learning, using pre-trained models on vast datasets to extract contextual and semantic meanings beyond simple word frequency counts. The multidimensional text analysis system dynamically adapts to the computational resources available and user needs, choosing between these methods to balance accuracy and efficiency.

From a computational point of view, the system is optimized to handle long texts efficiently by splitting long texts into manageable chunks to accommodate models with input length limitations. Chunks are processed in parallel to reduce computation time and results from each chunk are combined to produce overall dimension scores. Robust error handling ensures the system operates smoothly by capturing and handling exceptions without crashing the system. Using fallback mechanisms and defaulting to the standard method if the advanced method fails assures that the user still gets a result, thus saving time. Using input validation by

checking for empty texts, unsupported characters, or other anomalies add to the efforts of providing robustness to the model.

5. Understanding Dimensions

In this section, we will explain a few dimensions in order to create a general understanding about the reasoning behind our imagining of the system. We think that the term ‘dimension’ best describes what we are trying to achieve, since it encompasses a more general sense regarding the way of considering something. So, because our system tries to provide a ‘bird’s eye view’ regarding a text, measuring ‘dimensions’ of the text seemed appropriate.

The affective dimension means to capture the emotional flavour of a text. It indicates the extent to which the text elicits emotional responses, whether positive or negative. It is a general measure of the text's emotional tone, beyond specific sentiments like positivity or negativity. The affective dimension draws from psychology and linguistics, particularly theories of emotional engagement. It recognizes that affect plays a crucial role in how texts are perceived and understood, influencing cognition, persuasion, and overall reader engagement (Forgas et al., 1995).

The standard method calculates affective score using NLTK's opinion lexicon (Hu and Liu, 2004). It detects positive and negative emotions in the text using pre-defined word lists (affective lexicons) to measure the presence of emotional words. This method looks at the frequency of positive and negative emotional words and provides a score based on the emotional intensity of the text. Words like *happy*, *love*, *angry*, *fear*, *sad*, etc., are counted to determine the level of emotional content.

The input text is converted to lowercase and then tokenized into individual words using `word_tokenize`. Afterwards, it counts how many tokens appear in the `positive_words` set and how many in the `negative_words` set and computes the `total_words` from the number of tokens. The raw score is calculated as in (1).

$$\text{Affective_score} = ((\text{pos_count} - \text{neg_count}) / \text{total_words}) \times 100 \quad (1)$$

This score is normalized between 0 and 100. If there are no tokens (i.e., `total_words == 0`), the method returns a score of 0 to avoid division by zero. The score is clamped to the range 0 to 100. This ensures that even if the computed value is negative or above 100, the final score will be within the defined bounds. A text with a high number of affective words (relative to its total word count) will have a high score, while a text with no affective words will score 0.

The advanced method employs Sentiment Analysis using DistilBERT (Devlin et al., 2019) (distilbert base uncased finetuned sst 2 english) to capture the

affective undertone of the entire text. Instead of relying mainly on word frequency, the advanced method analyses contextual embeddings to understand the emotional depth of sentences and paragraphs. The model generates an overall emotional score based on how it interprets the text emotional undertone, rather than relying purely on word frequency.

The animate dimension measures the presence of animate entities, such as people or animals, in a text. High scores in this dimension suggest that living beings have importance, either through storytelling or descriptions of human or animal activity. The animate dimension stems from discourse analysis and cognitive linguistics, which recognize that texts focusing on animate entities often create more engagement and emotional involvement due to humans' natural affinity for living beings (Bonin et al., 2014; New et al., 2007).

The standard method relies on predefined lists of animate nouns (e.g., man, woman, dog, child) to measure the presence of living beings in the text. A simple frequency count of animate words relative to the total word count, includes words like person, dog, family, employee, teacher, cat, etc. The raw score is the number of animate references (people, animals) divided by the total number of words, as in (2).

$$\text{Animate Score} = (\text{Number of Animate Words} / \text{Total Words}) \times 100 \quad (2)$$

The advanced method utilizes Named Entity Recognition using BERT (dbmdz/bert large-cased finetuned conll03 english) (for person entities) to conduct named entity recognition (NER) to identify animate entities dynamically. But here lies a problem: the actual entity recognition models are not trained to recognize animals, only people. Since our dimension means to express how animate is the text, other animate entities besides just people should be considered. So, the advanced method also uses a predefined comprehensive animals list in order to complement the BERT model, thus combining existing methods in order to obtain a more valuable result. The model identifies entities like humans, animals, or groups and normalizes the score based on entity occurrence relative to the total text. Other entity-based dimensions, like Inanimate, Personality, or Character, would follow similar normalization strategies, focusing on different categories of entities. This dimension is an example of why the system needs to be customizable in order for each user to obtain what interests her by adjusting the way dimensions are calculated.

The cognitive dimension assesses the intellectual and logical flavour of the text. It tries to capture how much the text emphasizes reasoning, problem solving, or other cognitive processes. Rooted in cognitive linguistics and psychology, this dimension aligns with theories about how individuals process information and make decisions based on logical reasoning rather than emotions. It is not in

contrast to the affective dimension but rather coexists with it, adding detail to the whole picture.

The standard method uses comprehensive sets of cognitive words, phrases, mental process terms and intellectual adjectives to calculate their ratios to the total number of words and then computes the cognitive score using a weighted formula (3) and scaling the final score to 0-100:

$$\text{Cognitive Score} = (\text{cognitive_word_ratio} * 0.3) + (\text{cognitive_phrase_ratio} * 0.3) + (\text{mental_process_term_ratio} * 0.2) + (\text{intellectual_adj_ratio} * 0.2) \quad (3)$$

The advanced method uses Zero-Shot Classification using BART-large-MNLI (facebook/bart large mnli) to understand the overall logical structure and cognitive emphasis of the text. The advanced method would recognize reasoning even when expressed indirectly, such as in meta-analyses or theoretical discussions. Also, to ensure the advanced method outperforms the standard method, it performs a weighted combination with heuristic counts using the standard method sets. These weights may be adjusted, but for our prototype we have used the model 70% and standard 30% weights. (4)

$$\text{Combined Cognitive Score} = (\text{average_cognitive} * 0.7) + (\text{standard_score} * 0.3) \quad (4)$$

As shown, for each dimension we tried to identify free tools that can offer generally reliable results for describing texts. The customization attribute of our system implies that the user is able to use whatever tool she finds appropriate for the job. Going forward, we envision a user interface that will allow the selection of different free tools available to calculate each dimension, or to even use multiple tools comparatively or combined, in order to make this tool accessible to users who are not necessarily literate regarding the technical and NLP aspects of this system.

6. Understanding Synergies

In our system, synergies refer to the common occurrence of high dimension scores, possibly yielding a deeper understanding of the text beyond what individual dimensions can offer. With dimensions already providing a general view of the text, synergies would further describe the text at a higher level, by stressing the associations between dimensions and maybe hint regarding certain common features of the text at the meta-level. The idea is that understanding how these dimensions correlate can provide a more nuanced interpretation of the text nature, tone, or intent. Synergies are detected by examining which dimensions score high and looking for logical or meaningful connections between them. The system sets thresholds (e.g., scores above 60% for relevant dimensions) and checks whether the text exhibits multiple high-scoring dimensions that align with

described synergies. The number of dimensions that can create a synergy is not a fixed concept, nor the weight of each dimension within a certain synergy and we are convinced that by using the system, more and more relevant correlations between dimensions will surface in time. For now, we have tried to identify a few synergies that are more likely to shed a different light on the text when activated. Although, some of the synergies can only be inferred from tangential literature, or for some there is no literature to specifically address the exact combinations of word types into the same text, we still think they might hold value by highlighting certain accents of the target text and serve as relevant examples for the description of the system as a whole. Further research of academic sources in the fields of text analysis, cognitive psychology, linguistics, and communication studies as well as user feedback and future analysis of texts are required to properly identify and prove specific synergies, but for this prototype we have identified a few possibly relevant synergies. We will describe some of them in what follows.

The presence of the *Affective-Cognitive Synergy* reflects that there is an interaction between emotional engagement and logical reasoning in the target text. It could reflect the idea that, in the target text, the emotional elements impact cognitive processing, such as decision-making and reasoning (Forgas, 1995).

The *Cognitive-Dynamic Synergy*: the presence of a lot of cognitive and dynamic words in the same text could indicate strategic planning or planned actions or a description of many actions, or could even mean nothing, but the simple information that a text is cognitive and dynamic before even reading it, might hold relevance for users (Biber, 1988; Claussenius-Kalman et al., 2021).

The *Social-Cognitive Synergy* refers to the interplay between social dynamics and cognitive processes, reflecting how social interactions and structures influence cognitive development and vice versa (Clark, 1996; Pickering and Garrod, 2013). This is often explored in sociolinguistics and discourse analysis studies.

The *Time-Cognitive Synergy* examines the relationship between temporal constructs and cognitive processes. It could be particularly useful for identifying texts that deal with changes over time or historical perspectives that affect cognitive framing or the decision-making processes (Schilperoord, 1996; Bublitz and Norrick, 2011).

The *Affective + Social + Individual synergy* refers to the combination of emotional appeal (Affective), social context (Social), and personal perspective (Individual) and could be important in texts that reflect personal experiences in social settings. This could appear in autobiographies, testimonials, or opinion pieces where the writer's emotions are tied to both their individual identity and broader social dynamics (Ambrasat and von Scheve, 2022; Cheng, 2024; De Stefani and De Marco, 2019).

The *Cognitive + Time + Dynamic Synergy* should appear in texts where logical analysis (Cognitive) is applied to events unfolding over time (Time) with a focus on change (Dynamic). It could be relevant for understanding processes that evolve or develop, such as scientific research or historical analysis (Robinson et al., 2012).

The *Positive + Intentionality + Novelty Synergy* refers to texts with new, innovative ideas (Novelty) framed in a positive light (Positive) with clear intentions (Intentionality) are often visionary in nature. This synergy could be important for proposals, strategic plans, or futuristic literature, where the goal is to inspire progress (Kshenovskaya, 2020; Popova, 2014).

The synergies allow us to explore how different dimensions like emotion, logic, time, social context, and data specificity interact to create a more nuanced and effective message.

7. Visualization via Radar Chart

Visualizing the multidimensional analysis helps in quickly understanding the overall characteristics of the text, radar charts being particularly effective for this purpose. A radar chart displays multivariate data in a two-dimensional chart, where each dimension represents one axis radiating from the centre and each of the 26 dimensions is represented as an axis. The score for each dimension is plotted along its corresponding axis and the points are connected to form a polygon, visually representing the text profile. The radar chart can be interpreted considering the shape of the polygon to see the dominant dimensions, the area covered to intuit the overall intensity of characteristics, or the symmetry suggesting balanced dimension scores. For multiple files processing, the app also integrates all the results into the same chart for ease of comparison (see Figure 2). The modularization and the usage of widely and freely available tools make the visualization section of the app also easily customizable.

The third chart in Figure 2 shows the results for *Cicero's first speech against Catiline, delivered in the Roman Senate* (Cicero, 63 BCE), where the app correctly identifies the affective, negative and specific accents of the text by attributing high scores.

A MULTIDIMENSIONAL APPROACH FOR TEXT ANALYSIS



Figure 2: Single and multiple url/file representation using the radar chart

8. Implementation and Evaluation

For the implementation of the concept, we have developed a functional prototype of the app and we have utilized many publicly available, free texts and books (Project Gutenberg, n.d.), as well as AI generated texts for training and evaluation. The AI generated texts were used for the sole purpose of testing specific dimensions and synergies in order to obtain reliability for their calculation methods and this endeavour still represents a work in progress for every dimension. One avenue for further development is represented by creating tailored performance indicators/metrics for each module of the app. We know that the best performance indicator remains the user and we hope that going forward we will be able to get valuable feed-back in order to significantly improve the dimension and synergy calculation methods.

Analysis Report for All Inputs

General Analysis Report for All Inputs:

Article 2: Positive, Long-Term, and Social.docx:

Article 2: Positive, Long-Term, and Social.docx Analysis Report:

 Affective: 100.00
 Animate: 0.00
 Commonality: 0.00
 Cognitive: 70.02
 Directness: 0.00
 Dynamic: 0.00
 Formality: 0.00
 Generic: 0.00
 Inanimate: 0.00
 Individual: 0.00
 Informality: 0.00
 Intentionality: 0.00
 LongTerm: 0.00
 Negative: 0.14
 Novelty: 70.08
 Objectivity: 0.00
 Place: 0.00
 Politeness: 100.00
 Positive: 0.18
 Qualitative: 0.00
 Quantitative: 0.33
 ShortTerm: 70.07
 Social: 70.49
 Specific: 76.10
 Static: 70.05
 Time: 70.14

Synergy Analysis:

The high Affective and Cognitive scores suggest that emotional engagement is influencing logical reasoning in this text. This synergy is critical in persuasive writing, where emotional appeal is balanced with well-structured argumentation. Academic studies have shown that emotional elements can enhance cognitive processing, especially in contexts like advertising, political speeches, and motivational talks. (Oxford Academic: <https://academic.oup.com>) (SpringerLink: <https://link.springer.com>).

Figure 3: Excerpt from the comprehensive report

The app produces a comprehensive, downloadable report, including the scores for each of the dimensions, the interpretation of the identified synergies if any (see Figure 3), the overall interpretation showing the top 3 strongest dimensions and the top 3 weakest dimensions, the analysis method used and the reason behind zero scores - whether the text lacks the characteristics for that dimension or there was an error computing, the radar chart for the single file and the radar chart for all the files if the file in question was analysed along with others (see Figure 2).

9. Discussion

The multidimensional framework offers a comprehensive analysis across multiple dimensions and introduces the concept of predefined synergies to reveal compound effects, allowing flexibility based on resource availability. As it stands, some of the proposed dimensions and synergies are grounded in established theories from linguistics, psychology, and communication studies, some are inferred from interdisciplinary, complementary published papers, while others are intuition-based. As scientific researches advance, further dimensions and synergies will be discovered and proved, enriching the system and producing fresh insights. The system provides clear visualizations and comprehensive reports, enhancing understanding of the target text, the possibility of adding new dimensions and adjusting the existing ones, as well as the local implementation, providing adaptability and confidentiality.

Although the high-performance mode may not be feasible in all environments and the currently focus on English limits multilingual usage for now, the visualization using radar graphs bridges these limitations, providing a universal way to understand texts.

As long as we ensure that the models do not perpetuate biases from training data and we safeguard sensitive information when analysing personal or confidential texts, we believe this system could add valuable new perspectives regarding text analysis.

We envision several practical uses for this system. It could serve as an initial sorting tool for AI customer relations agents in producing ‘to the point’ replies. Also, it could power an email sorting system, as well as a snippet attached to the general section of the email app, to display the multidimensional graph with the title of the email, so that the user knows what to expect from reading it. It could aid the search engines to better interpret user requests and selecting which results are more relevant. In security, counter terrorism and public order activities, the system could be used to analyse and sort sensitive intercepted conversations. Psychologists could use it to obtain an extra point of view regarding their clients by running the system on all the client history of therapeutic conversations.

We think that new ways to work this system will be found by the users over the time.

10. Conclusions

Our multidimensional approach to text analysis offers a nuanced and interpretable framework, capturing the multifaceted nature of language. By integrating synergies between dimensions, the system provides the possibility of obtaining deeper insights that enhance applications in discourse analysis, predictive modelling, and content evaluation. As future work, we intend to improve our system from extending it in order to support multiple languages, upgrading the interface and perfecting the way each dimension is calculated, to implementing machine learning methods for automatic identification of new synergies or optimizing the algorithms and leveraging hardware acceleration for faster performance. The system versatility makes it suitable for various domains, like **academic research** for linguistics, psychology, and communication studies, in **corporate analysis** for market research, brand monitoring, and customer feedback analysis, in **education** for assisting in writing instruction and assessment or in **healthcare** for analysing patient narratives for psychological assessment, to name a few. An overall text analysis tool could open new possibilities if used as a component for even larger systems of interpretation.

References

- Ambrasat, J., and von Scheve, C. (2022). Affective meanings and social relations: identities and positions in the social space. *Emotions and Society*, 4(2), 161-180. <https://doi.org/10.1332/263169021X16340311969719>
- Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press, Cambridge, UK, <https://doi.org/10.1017/CBO9780511621024>
- Bonin, P., Gelin, M. and Bugaiska, A. (2014). Animates are better remembered than inanimates: further evidence from word and picture stimuli. *Memory and Cognition* 42, p. 370–382, <https://doi.org/10.3758/s13421-013-0368-8>
- Brown, P. and Levinson, S. C. (1987). *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK. ISBN: 978-0521576407, <https://doi.org/10.1017/CBO9780511813085>
- Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. <https://doi.org/10.48550/arXiv.2005.14165>
- Bublitz, W. and Norrick, N. R. (2011). *Foundations of Pragmatics*, De Gruyter. ISBN: 9783110214260, DOI: 10.1515/9783110214260
- Cheng K.C.-L. (2024). Self-Narrative, Affective Identification, and Personal Well-Being. *Journal of the American Philosophical Association*. 10(1), p. 79-95. DOI:10.1017/apa.2022.39
- Cicero's first speech against Catiline, delivered in the Roman Senate (63 BCE), translated by Charles Duke Yonge, Retrieved April 16, 2025, from <https://www.sjsu.edu/people/cynthia.rostankowski/courses/HUM1AF14/s3/Lecture-26-Cicero-and-Caesar-Reading.pdf>
- Clark, H. H. (1996). *Using Language*. Cambridge University Press. ISBN: 978-0521578647, <https://doi.org/10.1017/CBO9780511620539>
- Claussenius-Kalman H., Hernandez A.E. and Li P. (2021). Expertise, ecosystem, and emergentism: Dynamic developmental bilingualism. *Brain and Language*, Vol. 222; article number 105013, Epub 2021 Sep 11. PMID: 34520977. DOI: 10.1016/j.bandl.2021.105013
- De Stefani E. and De Marco D. (2019) Language, Gesture, and Emotional Communication: An Embodied View of Social Interaction. *Frontiers in Psychology* 10:2063, 1-8, DOI: 10.3389/fpsyg.2019.02063
- Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics – HLT*, 4171-4186. DOI:10.18653/v1/N19-1423
- Forgas, J. P. (1995). Mood and judgment: The affect infusion model (AIM). *Psychological Bulletin*, 117(1), 39-66. https://www.researchgate.net/profile/Joseph-Forgas/publication/15328091_Mood_and_judgment_The_Affect_Infusion_Model_AI_M/links/00b7d5159f3a346118000000/Mood-and-judgment-The-Affect-Infusion-Model-AIM.pdf
- Graesser, A. C., McNamara, D. S. and Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5), 223-234. <https://doi.org/10.3102/0013189X11413260>

- Heylighen, F. and Dewaele, J.-M. (1999). Formality of language: Definition, measurement, and behavioral determinants. Internal report, Brussels: Center “Leo Apostel”, Vrije Universiteit Brussel. <http://pespmc1.vub.ac.be/Papers/Formality.pdf>
- Hu, M. and Liu, B. (2004). Opinion Lexicon information. Department of Computer Science, University of Illinois at Chicago. Retrieved April 16, 2025 from https://www.nltk.org/api/nltk.corpus.reader.opinion_lexicon.html
- Kshenovskaya, U. (2020). Linguistic Creativity: Cognitive And Communicative Aspects. In N. L. Amiryannova (Ed.) Word, Utterance, Text: Cognitive, Pragmatic and Cultural Aspects, vol 86. *European Proceedings of Social and Behavioural Sciences*, 845-853, <https://doi.org/10.15405/epsbs.2020.08.98>
- New, J., Cosmides, L. and Tooby, J. (2007). Category-Specific Attention for Animals Reflects Ancestral Priorities, Not Expertise. *Proceedings of the National Academy of Sciences*, 104(42), 16598-16603. <https://doi.org/10.1073/pnas.0703913104>
- Pennebaker, J. W., Booth, R. J. and Francis, M. E. (2015). Linguistic Inquiry and Word Count. 1-24. Pennebaker Conglomerates. https://www.researchgate.net/publication/337731895_Linguistic_Inquiry_and_Word_Count_LIWC2015
- Pickering, M. J. and Garrod, S. (2013). An Integrated Theory of Language Production and Comprehension. *Behavioral and Brain Sciences*, 36(4), 329-347. DOI: 10.1017/S0140525X12001495
- Popova Y.B. (2014) Narrativity and enaction: the social nature of literary narrative understanding. *Front. Psychol.* 5:895. 1-14. <https://doi.org/10.3389/fpsyg.2014.00895>
- Project Gutenberg. (n.d.). Project Gutenberg: Free ebooks. <https://www.gutenberg.org/>
- Robinson C.W., Best C.A., Deng W.S. and Sloutsky V.M. (2012). The role of words in cognitive tasks: what, when, and how? *Front. Psychology* 3:95. 1-8. <https://doi.org/10.3389/fpsyg.2012.00095>
- Schilperoord, J. (1996). It's about time: Temporal Aspects of Cognitive Processes in Text Production. *Utrecht Studies in Language and Communication*, Editions Rodopi B.V., Utrecht, The Netherlands. ISBN: 90-5183-947-2

ENHANCING PATIENT COMPREHENSION OF MEDICAL REPORTS THROUGH LLM-BASED LANGUAGE SIMPLIFICATION: THE CASE OF MEDX

VALENTIN NAGACEVSCHI¹, TUDOR P. TOMA²

¹*MyTechMed Ltd. London, UK*

vali.nagacevschi@gmail.com

²*University Hospital Lewisham, London, UK*

ttoma@doctors.org.uk

Abstract

MedX introduces a novel interface utilising large language models (LLMs) to translate complex medical terminology into simple language comprehensible to individuals without medical training, equivalent to a 16-year-old's understanding. This paper presents the design and preliminary testing of MedX, which uses generative AI and natural language processing techniques to render blood tests, pathology and radiology results, and other medical reports into easily understandable language. The development process involved collaborative prompt refinement between experienced developers and clinicians to ensure the medical accuracy and readability of outputs. MedX prioritises user privacy with a strict no-data-storage policy and offers free document uploads for analysis. Initial user feedback was collected for the first 58 reports, focusing on usability, utility, and user satisfaction. Clinical review of these reports confirmed the translations' accuracy, appropriateness, and empathetic tone. Preliminary results suggest that MedX effectively makes medical reports accessible and useful to the public. Future work will explore broader acceptance among users and healthcare professionals and expand MedX's features to include online consultations and health monitoring, aiming to position it as a comprehensive digital health tool. This study highlights the potential of LLM applications in bridging the communication gap between medical professionals and patients, thus enhancing health literacy and decision-making.

1. Introduction

The ability of patients to understand medical reports is essential for informed decision-making and effective management of their health. However, reading these reports can be a formidable challenge for many, influenced by an array of factors that impair their ability to comprehend essential health information.

Medical reports often employ technical jargon that is unfamiliar to non-specialists. For instance, a national survey in Japan highlighted that over a third of respondents had difficulties with medical terminology, with specific terms like “Kakutan Saibo-shin” (sputum cytology) proving particularly troublesome (Tokuda et al., 2008). Furthermore, pathology reports, although crucial, are not always understood well; terms such as “malignant” and “benign” are recognized by the majority, yet more complex terms like “high grade” are correctly understood by only a small fraction (Verosky et al., 2022). This disparity in comprehension emphasises the gap between the language used by healthcare professionals and the linguistic accessibility required by patients.

Health literacy also plays a pivotal role in patient understanding. Research indicates that patients with limited health literacy are significantly less likely to receive patient-centred communication, which is important for making informed health decisions (Wynia et al., 2010). Moreover, demographic factors such as race and educational attainment can exacerbate disparities in health literacy, as evidenced by studies showing that non-White patients and those with lower educational levels often have a poorer understanding of medical terms (Fisher et al., 2024).

Language barriers further complicate communication between healthcare providers and patients, particularly in multicultural settings. In Japan, language discordance has been shown to significantly impact patient comprehension, necessitating the use of medical interpreters (Oonishi et al. 2023). Similarly, migrant populations in Australia face challenges in understanding medical information, with language and cultural differences influencing the degree of comprehension (Hyatt et al., 2017).

The format and presentation of medical reports are additional hurdles. Traditional pathology reports are not typically designed with the patient perspective in mind, leading to confusion and misinterpretation (Austin et al. 2020). The advent of electronic health records has offered new avenues for patient engagement through portals like My Kanta in Finland, yet challenges persist in how medical information is structured and presented (Kujala et al., 2022).

Emotional and psychological factors also influence how patients interact with and interpret medical reports. Anxiety induced by medical terminology can lead to miscommunication and increased distress (Rogers et al., 2023). The Cures Act’s mandate for the immediate release of health information has resulted in varied comprehension levels among patients, highlighting the need for clearer explanations and more navigable information within medical reports (O’Brien et al., 2024).

This paper introduces MedX, a novel interface that employs large language models (LLMs) to simplify medical jargon into plain language accessible to the

public, aiming for comprehension at the level of a 16-year-old. Using a specific protocol for accessing generative AI and natural language processing techniques, MedX transforms blood tests, pathology and radiology results, and other medical documents into easily understandable text.

This is an example of a project that not only promises to bridge the communication gap between healthcare providers and patients but also to empower patients by making health information more accessible and actionable.

The development, design, and preliminary testing phases of MedX are detailed herein.

2. Review of Literature

The utilisation of LLMs in natural language processing (NLP) tasks has garnered substantial attention in both academic research and practical applications. Models such as BERT (Devlin et al., 2018) and GPT-3 (Brown et al., 2020) have set new benchmarks in various language tasks, including machine translation, text summarization, and question-answering systems. Their ability to process and generate contextually relevant, human-like text has opened possibilities for applications across diverse domains, notably in healthcare.

In the healthcare sector, NLP has been instrumental in tasks such as electronic health record (EHR) analysis (Hossain et al., 2023), predictive diagnostics, and patient communication tools. For instance, it has been demonstrated (Li et al., 2023) the efficacy of NLP models in extracting clinically relevant information from unstructured medical texts, thereby aiding in clinical decision support systems.

Despite these advancements, challenges persist in the application of LLMs to medical language simplification. Medical terminology is inherently complex, context-dependent, and often contains nuances that are critical for accurate understanding. While LLMs exhibit proficiency in general language tasks, they may lack the domain-specific knowledge required for precise medical interpretations (Nazi et al., 2024). This limitation emphasises the necessity for collaborative efforts between AI developers and medical professionals to ensure that AI-generated content maintains clinical accuracy and relevance (Zhang et al., 2023).

Existing applications like HealthLynk and MyHealthPal have made steps in improving patient health literacy through AI-driven simplification of healthcare information. However, these tools often focus on specific health conditions or types of medical documents, limiting their applicability. Moreover, concerns regarding data privacy and security remain prevalent, necessitating solutions that not only simplify information but also safeguard patient data in compliance with

regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA).

MedX aims to build upon this foundation by providing a simple, broad-spectrum, solution capable of handling multiple report formats while ensuring user privacy and data security. By integrating medical expertise into the development process, MedX aims to overcome the limitations of existing models and offer a reliable tool for enhancing patient comprehension.

3. Methods

3.1. Development Process

The primary objective was to create and test a simple, user-centric, application capable of processing various types of medical reports and providing simplified, accurate explanations without compromising patient privacy.

The development of MedX involved a methodical approach to ensure the medical accuracy and readability of the translated outputs.

This process was a collaborative effort between an experienced software developer (VN) and a licensed clinician (TPT). While TPT is a clinician licensed by the NHS, it is important to note that the development and testing of the MedX application were conducted outside of UK NHS institutions.

Data Preparation

An important component of the project was the selection of a dataset comprising 58 anonymized medical reports. While the sample size is modest and not exhaustive, it provided a diverse range of medical documents for initial testing and development. Each report was sourced from a different individual, ensuring that all records belonged to unique patients with no returning users. This approach was taken to capture a wide variety of medical terminology and report structures.

The distribution of the report types among the data was as follows:

Blood Tests: 25 reports (e.g., complete blood counts, metabolic panels)

Radiology Reports: 15 reports (e.g., X-rays, CT scans, MRI findings)

Pathology Reports: 10 reports (e.g., biopsy results, histopathology findings)

Echography Reports: 5 reports (e.g., ultrasound examinations)

Genetic Analysis Reports: 3 reports

These reports were collected following a short campaign on Facebook targeting patients in Romania. Participants agreed to contribute their medical documents by signing a terms and conditions agreement specific to the pilot application. The diversity of the documents aimed to ensure that the application could handle various medical terminologies and formats.

This activity was performed independently of the UK NHS and its associated facilities.

3.2. Model Fine-Tuning and Prompt Engineering

To facilitate the transformation of complex medical terminology into plain language, MedX utilises a specific protocol designed for effective interaction with LLMs. Through an iterative process, specific prompts were designed and refined to guide the model in generating clear, accurate, and empathetic explanations. The iterations were performed on a small sample of 10 synthetic reports created to mimic typical medical documents, allowing for focused testing without compromising patient data.

Examples of Prompt Iterations

Initial prompt:

Translate the following medical report into simple language understandable by a 16-year-old.

Outcome: The LLM simplified the text but occasionally omitted critical medical details and lacked an empathetic tone.

Refined prompt (Iteration 1):

Please explain the key points of this medical report in plain language suitable for a teenager; ensuring all important details are included.

Outcome: Improved inclusion of important details but still somewhat technical and impersonal.

Refined prompt (Iteration 2):

As a friendly doctor, explain this medical report to a 16-year-old patient, using simple language and an empathetic tone. Make sure to cover all essential findings without using medical jargon.

Outcome: Better simplicity and tone, with more emphasis on empathy, but occasional use of medical jargon persisted.

Final prompt:

You are a compassionate medical professional explaining the following medical report to a 16-year-old patient in plain language. Avoid medical jargon, simplify complex terms, and maintain an empathetic and reassuring tone. Ensure that all key findings and their implications are clearly described.

Outcome: The LLM provided clear, accurate explanations with an appropriate level of detail and an empathetic tone suitable for the target audience.

The iterative refinement focused on:

Clarity and simplicity: Encouraging the LLM to use everyday language and explain medical terms when unavoidable.

Empathetic tone: Instructing the model to adopt a compassionate and reassuring manner to address potential patient anxieties.

Accuracy: Ensuring that essential medical details were preserved and accurately conveyed.

Comprehensiveness: Guiding the model to cover all key findings without overwhelming the reader with unnecessary information.

Feedback from the clinician was instrumental in each iteration, highlighting areas where the LLM’s responses needed adjustment. The developer used this feedback to modify the prompts, aiming to balance simplicity with completeness and empathy.

The selection of the LLM took into consideration multidimensional criteria such as accuracy, readability and comprehension, and tone and empathy.

Tests were conducted among the three top LLMs publicly available, namely OpenAI’s GPT-4o, Google’s Gemini 1.5, and Meta’s LLama 3.2. OpenAI’s GPT-4o was selected for its advanced language capabilities and adaptability, as it scored the highest across all evaluated dimensions.

3.3. Privacy and Security

Given the sensitive nature of medical information, MedX was designed with strict privacy and security protocols. However, we recognize that data transmission over the web has inherent vulnerabilities, and the submission of medical data to external systems like LLMs poses additional risks.

To mitigate these vulnerabilities, the following measures were implemented:

End-to-end encryption: All data transmissions between the user’s device and the server are secured using HTTPS with TLS encryption, preventing unauthorised interception and access during data transfer.

No data storage policy: The application operates under a strict no-data-storage policy, where uploaded reports are processed in real-time memory and are not stored on the server or any databases post-processing. This reduces the risk of data breaches from stored files.

Secure processing environment: The server environment is secured with firewall protections and regular security updates to prevent unauthorised access.

Data handling with LLMs: Recognizing that LLMs like OpenAI’s GPT-4o may collect input data for training purposes, we applied for a Business Associate Agreement with OpenAI and obtained the **Zero Data Retention** status, thus guaranteeing the data protection.

User consent and transparency: Users are informed about the data processing methods and potential risks through clear terms and conditions before uploading their documents. They must provide informed consent acknowledging these terms.

Compliance with regulations: The application underwent a comprehensive internal audit to ensure adherence to GDPR and HIPAA regulations, including the implementation of user consent mechanisms, transparency in data handling practices, and the right to request data deletion.

Despite these measures, we acknowledge that submitting medical data over the internet to third-party services like LLMs cannot be entirely risk-free. Users are advised of these risks, and future development will explore additional security measures, such as on-device processing or the use of private LLMs that do not transmit data externally.

3.4. User Interface and Experience

The user interface of MedX was designed to be intuitive and accessible, requiring minimal user input.

Users can easily upload medical reports in various formats, including PDF, DOCX, and JPEG.

Upon upload, the application generates simplified explanations almost instantaneously, enhancing the user experience through efficiency and convenience.

The design principles emphasised simplicity and ease of navigation, avoiding complex menus or settings that could overwhelm users.

There is no requirement for account creation or the disclosure of personal information, aligning with the application's privacy commitments.

Accessibility considerations were integrated into the design to accommodate individuals with disabilities. The application adheres to the Web Content Accessibility Guidelines (WCAG), ensuring that features such as text-to-speech compatibility, keyboard navigation, and appropriate colour contrasts are available to users who need them.

This was intended to encourage widespread usage and facilitate easy access to the service, promoting better understanding of medical information across diverse populations.

The first screenshot displays the homepage, featuring a user-friendly design with clear navigation options and prompt users to enter their lab test results. The 2nd screenshot illustrates the How to Use and FAQ page, to help users understand the application. The 3rd screenshot asks the users to read and agree with the terms and conditions. The 4th screenshot showcases the AI-generated explanation of a specific lab test, highlighting the simplified language used to make complex medical information more accessible to patients. Finally, the 5th screenshot presents the feedback section, offering users the opportunity to provide valuable input, to be used to further enhance application performance.

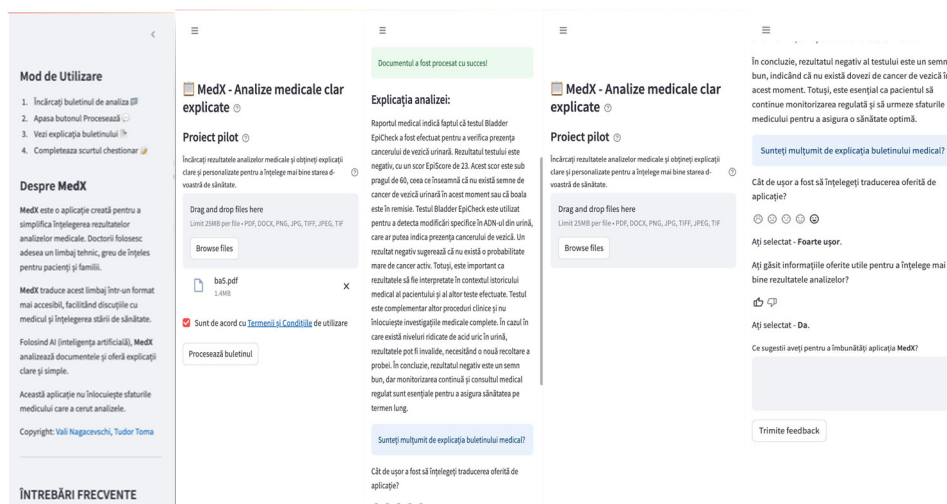


Figure 1: Screenshots of the MedX application mobile interface.

3.5. Preliminary Testing and User Feedback Collection

This report describes the preliminary testing of MedX which was conducted through the analysis of the first 58 medical reports uploaded by users.

The testing focused on three primary dimensions: translation accuracy and appropriateness, usability and utility, while the user had also an option for free text comments.

A clinical review was conducted to assess the accuracy and appropriateness of the translations provided by MedX. This review involved a clinician (TPT) who examined all translated reports against the original medical documents.

Usability assessed how easily users could navigate the interface and upload documents for translation.

Utility examined the effectiveness of the MedX translations in aiding users' understanding of their medical reports.

4. Results

The preliminary evaluation of MedX revealed the following outcomes across several metrics.

4.1. Translation Accuracy and Appropriateness

The clinical review of the outcomes confirmed that the translations accurately reflected the medical content while simplifying complex terminology into more comprehensible language.

The review also noted the appropriateness of the language used, ensuring that it was suitable for a general audience without compromising the integrity of the medical information. Importantly, the translations maintained an empathetic tone, which is critical in medical communications, helping to alleviate potential anxieties that patients might experience when reading about their health conditions.

Furthermore, MedX demonstrated robust capability in distinguishing relevant medical documents from non-relevant submissions. When users uploaded documents that were not human medical reports, the system effectively identified and declined to translate these. One instance involved a user who uploaded a veterinary report; the system appropriately recognized this as outside its specified scope and rejected the translation. This feature ensures that MedX maintains its focus on providing accurate and relevant medical translations for human healthcare documents.

4.2. Utility and Usability (or Readability and Comprehension)

Not all users provided feedback on utility and usability. Feedback was received from 38 users and was systematically analysed to identify common patterns and areas for improvement.

Users reported that the interface was intuitive and user-friendly, facilitating easy uploads and swift receipt of translated reports.

27 users expressed that the translated documents significantly aided their understanding of their medical conditions and treatment options, reflecting the high utility of the service.

Moreover, the accessibility features, such as the free document upload service, were particularly appreciated, making MedX a viable option for a wide range of users, irrespective of their financial or technological capabilities.

4.3. Influence of Demographic Factors

Demographic data such as age, race, education level, and migrant status were not systematically collected during this preliminary testing phase. Therefore, an analysis of how these factors influenced user perception of the system's performance is limited.

However, anecdotal feedback from users suggested that the simplified explanations were generally helpful across different educational backgrounds. Since all participants were Romanian patients who self-selected to participate via Facebook, the sample may not fully represent the diversity of potential users, particularly those with varying levels of health literacy or from different cultural or linguistic backgrounds.

5. Discussion

The preliminary results from the deployment of MedX indicate that the system is able to successfully translate medical reports into language that is accurate, appropriate, and empathetic, thereby making complex medical information accessible and useful to the general public.

These outcomes also validate the effectiveness of the collaborative development process involving developers and clinicians.

Other projects have explored the potential of Large Language Models (LLMs) like ChatGPT for various healthcare applications, notably for simplifying complex medical terminology to enhance comprehension among non-specialists.

Radiology Report Simplification. A study assessed four LLMs (ChatGPT-3.5, ChatGPT-4, Bard, and Bing) on their ability to simplify radiology report impressions from the Medical Information Mart for Intensive Care (MIMIC)-IV database, which includes 750 anonymized reports. The LLMs were tasked to adapt these reports to various comprehension levels, including that akin to a 7th-grade student, roughly equivalent to a 16-year-old's understanding. Methodology: The models received three types of prompts: "Simplify this radiology report," "I am a patient. Simplify this radiology report," and "Simplify this radiology report at the 7th grade level." Results: All models successfully simplified the radiology report impressions across the given prompts, showcasing their ability to render complex medical information more accessible (Doshi et al. 2024).

General Medical Information. A systematic review explored the application of LLMs like ChatGPT in various healthcare settings, including summarization and medical knowledge inquiry. The review acknowledged the promising outcomes of conversational LLMs in delivering medical information with relatively high accuracy, though it noted limitations in handling complex health-related tasks that demand specialised domain knowledge (Wang et al. 2024).

Medical Education. Research into the use of LLMs within medical education highlighted their potential to enrich learning experiences by personalising content and simplifying complex medical concepts. While this study primarily targeted medical students, it emphasises the broader applicability of LLMs in breaking down sophisticated medical information into more comprehensible formats (Lucas et al. 2024).

5.1. Challenges and Considerations

Accuracy and Reliability. Despite the potential of LLMs to simplify medical terminology, their performance can vary, particularly in complex medical scenarios. This variability highlights the importance of rigorous validation and

continuous oversight when employing these models in clinical environments (Wang et al., 2024; Sorin et al., 2024).

Ethical and Privacy Concerns. Utilising LLMs in healthcare raises significant ethical and privacy issues. It is therefore important to ensure that these models do not produce misleading or incorrect information, and that patient data is rigorously protected (Park et al., 2024; Preiksaitis et al., 2024). In our study we employed the experience of a senior clinician to verify the output information. However, once the application is validated it may not be necessary for a clinician to verify every report individually, although this remains to be demonstrated.

Our study suggests that LLMs like ChatGPT are capable of effectively simplifying medical jargon to levels understandable by a 16-year-old. Nevertheless, the challenges concerning accuracy, reliability, and ethical considerations must be carefully addressed to ensure the safe and effective use of these tools in healthcare. Future research should aim to refine these models further and develop standardised frameworks for evaluation to maximise their potential in making medical jargon comprehensible to a broader audience. By enhancing the capabilities of LLMs, healthcare providers can improve patient understanding and engagement, thereby potentially improving health outcomes.

5.2. Limitations of Our Study

This preliminary study, while promising, presents several limitations that must be acknowledged. Firstly, the size of the study is relatively small, encompassing only 58 self-selected patients. This sample size and the self-selection process may not adequately represent the diverse patient populations who might use MedX, potentially limiting the generalizability of the findings. Participants who are motivated to use such a service might also possess a higher level of health literacy or technological familiarity compared to the general population, which could skew the results towards more positive outcomes.

Secondly, the accuracy and appropriateness of the translations were evaluated by an in-house clinician who was involved in the development of MedX. This involvement raises potential biases in the evaluation process, as the clinician might inherently favour the performance of the application. Ideally, the evaluation of such a tool should be conducted by independent clinicians who are not affiliated with the project to ensure an impartial assessment. This would help validate the translations' accuracy and the system's utility in a more objective manner, enhancing the credibility of the study's outcomes.

Further research is required on a larger, more diverse sample and independent evaluations to fully ascertain the effectiveness and applicability of MedX in broader clinical and non-clinical settings.

6. Conclusions

The goal with projects like MedX is to further empower patients, providing them with the tools necessary to navigate their health information effectively. This study illustrates the potential of Large Language Models (LLMs) in bridging the communication gap between healthcare professionals and patients, thereby enhancing health literacy and informed decision-making. The preliminary assessment of MedX in simplifying medical reports to a level comprehensible by the public sets a promising basis for its future development.

Future work will focus on extending the data set and further testing with perhaps expanding the acceptance of MedX among a broader spectrum of users and healthcare professionals. Additionally, development may consider augmenting MedX's capabilities to include features such as online consultations and health monitoring. These enhancements may demonstrate MedX as an example of a comprehensive digital health tool that not only simplifies medical language but also facilitates a more interactive and integrated healthcare experience. This may contribute to better patient outcomes and a more informed patient populace, reinforcing the role of LLMs as pivotal in the evolution of patient-centred healthcare technologies.

The application can be accessed and evaluated at <https://medx.appcenter.ro>.

Author Contributions and Acknowledgements

The manuscript of this paper was collaboratively written by both authors, VN and TPT, who contributed equally to its creation. The references were carefully researched and selected from EvidenceHunt.com and PubMed (<https://pubmed.ncbi.nlm.nih.gov>). To enhance coherence and cohesion, parts of the text were edited and formatted using specific prompts in ChatGPT-4o. After these modifications, the draft was meticulously reviewed, edited, and corrected by both authors, who assume full responsibility for the content presented in this paper.

References

- Austin, Emily J., Theiss, Amanda L., and Oh, Rebecca C. (2020). "Help me figure this out": Qualitative explorations of patient experiences with cancer pathology reports. *Patient Education and Counselling*, 103(11), p. 2201 - 2207. <https://pubmed.ncbi.nlm.nih.gov/32800624/>
- Brown, Tom B., Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D., Dhariwal, Prafulla, ... and Amodei, Dario. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, p. 1877-1901. <https://arxiv.org/pdf/2005.14165>

- Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint. <https://arxiv.org/pdf/1810.04805>
- Doshi, Rohan, Sheikh, Muhammad, Johnson, Jonathon, Rajpurkar, Pranav, and Lundstrom, Cameron. (2024). Quantitative evaluation of large language models to streamline radiology report impressions: A multimodal retrospective analysis. *Radiology*. <https://pubmed.ncbi.nlm.nih.gov/38530171/>
- Fisher, Yael, Rose, Lora E., Narra, Sushma, Yousefi, Nazanin, and Cipriani, Nicole A. (2024). Evaluating patients' health literacy to improve thyroid pathology reports. *The Journal of Surgical Research*, 292, 143–150. <https://pubmed.ncbi.nlm.nih.gov/38701702/>
- Hossain, Ehsanul, Mahmud, Md Raihanul, Rahman, Md Saifur, and Sarker, Farhana. (2023). Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*, 155, 106649. <https://doi.org/10.1016/j.combiomed.2023.106649>
- Hyatt, Amber, Philip, Jennifer, Gold, Melissa, and Street, Annette. (2017). Communication challenges experienced by migrants with cancer: A comparison of migrant and English-speaking Australian-born cancer patients. *Health Expectations: An International Journal of Public Participation in Health Care and Health Policy*, 20(6), 1241–1250. <https://pubmed.ncbi.nlm.nih.gov/28261937/>
- Kujala, Sari, Hörhammer, Elisa, Heponiemi, Tarja, and Hilama, Pauliina. (2022). Patients' experiences of web-based access to electronic health records in Finland: Cross-sectional survey. *Journal of Medical Internet Research*, 24(6), e37901. <https://pubmed.ncbi.nlm.nih.gov/35666563/>
- Li, Ian, Dehghan, Ali, Karimzadeh, Maryam, and Hovy, Eduard. (2022). Neural natural language processing for unstructured data in electronic health records: A review. *Computer Science Review*, 46, 100511. <https://doi.org/10.1016/j.cosrev.2022.100511>
- Lucas, Henry C., Smith, Alexandra J., and Kramer, Brian S. (2024). A systematic review of large language models and their implications in medical education. *Medical Education*. <https://pubmed.ncbi.nlm.nih.gov/38639098/>
- Nazi, Zia Agha, and Peng, Wei. (2024). Large language models in healthcare and medical domain: A review. *Informatics*, 11(2), <https://doi.org/10.3390/informatics11020057>
- O'Brien, Ellen, Howard-McNatt, Marissa, and Sabel, Michael S. (2024). Interpretation of immediately released health information: Informing patient medical education in breast oncology. *American Journal of Surgery*. <https://pubmed.ncbi.nlm.nih.gov/39095250/>
- Oonishi, Akiko, Yamaguchi, Katsuhito, and Inoue, Satoshi. (2023). Relationship between patient-provider language discordance and the need for professional medical interpretation for international patients in Japan. *Cureus*, 15(10), e46756. <https://pubmed.ncbi.nlm.nih.gov/37965417/>
- Park, Yong Jin, Chen, Lin, Wang, Hua, and Liu, Fei. (2024). Assessing the research landscape and clinical utility of large language models: A scoping review. *BMC Medical Informatics and Decision Making*, 24, 22. <https://pubmed.ncbi.nlm.nih.gov/38475802/>

- Preiksaitis, Catherine, Yoo, Eric J., and Pavlik, Katherine. (2024). The role of large language models in transforming emergency medicine: Scoping review. *JMIR Medical Informatics*, 12, e54542. <https://pubmed.ncbi.nlm.nih.gov/38728687/>
- Rogers, Christopher, Tuck, Julianne, and Hughes, Sarah. (2023). Patient experience of imaging reports: A systematic literature review. *Ultrasound* (Leeds, England), 31(3), p. 186-196. <https://pubmed.ncbi.nlm.nih.gov/37538965/>
- Sorin, Veronica, Holmes, Courtney, and Davis, Matthew. (2024). Utilizing large language models in breast cancer management: Systematic review. *Journal of Cancer Research and Clinical Oncology*, 150(4), p. 913-925. <https://pubmed.ncbi.nlm.nih.gov/38504034/>
- Tokuda, Yasuharu, Goto, Eiji, and Otaki, Jun. (2008). The influence of medical jargon mixed with foreign terminology in the Japanese clinical environment. *Internal Medicine* (Tokyo, Japan), 47(12), 1083–1086. <https://pubmed.ncbi.nlm.nih.gov/18628581/>
- Verosky, Amanda, Kanchwala, Ayesha, and Allison, Kimberly. (2022). Patient comprehension of breast pathology report terminology: The need for patient-centered resources. *Surgery*, 172(4), 1308–1315. <https://pubmed.ncbi.nlm.nih.gov/35715235/>
- Wang, Lingling, Wei, Xinyue, Li, Jianbo, and Luo, Yu. (2024). A systematic review of ChatGPT and other conversational large language models in healthcare. medRxiv. <https://pubmed.ncbi.nlm.nih.gov/38712148/>
- Wynia, Matthew K., Osborn, Chandra Y., and Kaplan, Sally H. (2010). Health literacy and communication quality in health care organizations. *Journal of Health Communication*, 15 (Suppl. 2), 102–115. <https://pubmed.ncbi.nlm.nih.gov/20845197/>
- Zhang, Peng, and Kamel Boulos, Maged N. (2023). Generative AI in medicine and healthcare: Promises, opportunities and challenges. *Future Internet*, 15(9), 286. <https://doi.org/10.3390/fi15090286>

CHAPTER 4

LARGE LANGUAGE MODELS

EXPLORING LARGE LANGUAGE MODELS FOR TRANSLATING ROMANIAN COMPUTATIONAL PROBLEMS INTO ENGLISH

ADRIAN MARIUS DUMITRAN¹, ADRIAN-CĂTĂLIN BADEA²,
ȘTEFAN-GABRIEL MUSCALU³, ANGELA-LILIANA DUMITRAN⁴,
ȘTEFAN-COSMIN DĂSCĂLESCU⁵, RADU-SEBASTIAN AMARIE⁶

¹*University of Bucharest, Softbinator, Bucharest,*
marius.dumitran@unibuc.ro

²*University of Bucharest, UiPath, Bucharest,*
badeaadi1999@gmail.com

³*It Just Works Inc.*
stefan.gabriel.muscalu@gmail.com

⁴*University of Bucharest, Faculty of Foreign Languages and Literatures*
dumitranangela@gmail.com

⁵*QPillars, University of Bucharest*
stefdasca@gmail.com

⁶*It Just Works Inc., University of Bucharest*
raduamarie@gmail.com

Abstract

Recent studies have suggested that large language models (LLMs) underperform on mathematical and computer science tasks when these problems are translated from Romanian into English, compared to their original Romanian format. Accurate translation is critical for applications ranging from automatic translations in programming competitions to the creation of high-quality educational materials, as well as minimizing errors or fraud in human translations. This study shows that robust large language models (LLMs) can maintain or even enhance their performance in translating less common languages when given well-structured prompts. Our findings suggest that LLMs, with appropriate supervision, can be reliably used for the automatic translation of IOI (International Olympiad in Informatics)-style tasks. We evaluate several translation methods across multiple LLMs, including OpenRoLLM, Llama 3.1 8B, Llama 3.2 3B and GPT-4o, assessing their translation accuracy and performance stability through repeated runs. Additionally, we augment the OJI (Romanian County-Level Informatics Olympiad) Romanian dataset with accurate English translations, enhancing its utility for future LLM training and evaluation. Through detailed syntactic and semantic analyses, we confirm that with human oversight, LLMs can serve as a viable solution for multilingual problem-solving.

We also compare the translation quality of LLMs against human translators, as evaluated by a certified expert, underscoring the potential of LLMs in real-world scenarios.

Keywords: Automated Translation, Romanian to English Translation, Dataset Enhancement, Syntactic and Semantic Analysis, Translation Quality, LLM Training and Assessment

1. Introduction

Large language models (LLMs) have showcased remarkable capabilities across a wide range of natural language processing (NLP) tasks, including text generation, translation, code completion, and problem-solving in technical domains. Despite this progress, recent research has pointed to challenges in applying LLMs to structured domains, particularly in areas such as mathematics and computer science. Studies by Rae et al. (2021) indicate that while scaling LLMs improves performance across many tasks, structured tasks like mathematical problem-solving see smaller gains, suggesting that complexities in these domains are not fully captured or preserved. This discrepancy prompts further exploration of linguistic and computational factors that impact performance when translating such structured tasks across languages. Translating mathematical and computational tasks across languages presents challenges that go beyond typical linguistic differences. Technical problems require high levels of precision, and even minor translation errors – whether due to loss of mathematical context or subtle language ambiguities – can prevent humans from correctly understanding and solving these problems. When translations are flawed, they can hinder human problem-solvers by introducing confusion or misinterpretation, which is particularly impactful in high-stakes, multilingual environments.

In (Cosma et al., 2024b), the authors emphasize the need for dedicated resources beyond simple automatic translation, particularly for underrepresented languages like Romanian. Other researches such as Cosma et al. (2024a) argue for the necessity of developing code models for languages other than English, highlighting the current limitations of large language models in understanding and solving problems in non-English languages.

In (Dumitran et al., 2024), we delved into the performance of English large language models (LLMs) in solving competitive programming problems from the Romanian Informatics Olympiad at the county level. The study revealed significant variations in LLM performance across different grades and problem types, with GPT-4 showing strong performance.

The primary objective of this paper is to systematically analyse how translation-induced errors affect the ability of humans to solve technical problems when using LLMs as a translation aid. Focusing on Romanian-to-English

translations of IOI-style problems from the OJI dataset, we evaluate different translation strategies to identify the most reliable methods. Through repeated testing and performance analysis, we determine the best ways to ensure that translations retain the original meaning and clarity. In order to support human problem-solvers and to help LLM training, we enhance the OJI dataset with accurate English translations and propose an optimized prompt that improves translation accuracy, ensuring that humans can effectively solve the translated problems with the same or better understanding than from the original Romanian versions.

Although prior research exists on Romanian-English translation and the translation of mathematical problems, we have not identified any studies specifically addressing Machine Translation of computer science tasks between any language pairs. Furthermore, we are confident that no such work has been conducted for the Romanian-English language pair.

This paper is organized as follows: Section 2 outlines the methodology, detailing the translation approaches and the LLM evaluation process. Section 3 presents the results of our experiments, with a focus on translation accuracy and performance variability. In Section 4, we conduct a syntactic and semantic analysis of translation errors, and Section 5 introduces the optimal prompt and discusses the translation improvements it brings. Finally, Section 6 presents the conclusions.

2. Methodology

The methodology for this study was structured into several key stages, employing both quantitative and qualitative methods to thoroughly analyse the data.

1. In the initial stage, we chose 44 problem statements in Romanian from the same grade (15-16 years old), out of the 300 in "OJI" dataset introduced by Dumitran et al. (2024).
2. In the second stage, we computed a "Ro_score" for each problem, representing the highest score GPT-4o achieved after attempting to solve each problem five times in its original Romanian form.
3. The third stage involved translating the problems using a diverse set of LLMs (which will be detailed in 2.1), with variations in temperature settings for some models to assess their impact on translation quality.
4. In the fourth stage, GPT-4o was run five times with temperature 0.4 on each translated version of the problem, and scores were obtained for each task and each translation.

Importantly, we compared the scores achieved by GPT-4o on the original Romanian tasks (Ro_score) and the translated tasks, with GPT-4o acting as the

evaluator while the other LLMs handled only the translation process, without generating code.

A quantitative analysis was conducted to compare the judge scores across all LLM translations, with additional investigations into the effects of temperature settings and model size.

Next, we conducted an error analysis, where translations were reviewed and categorized by a linguist. Common issues identified included mistranslations, inappropriate language, untranslated content, inconsistent translation of algorithmic and computational terminology, and untranslated examples. Additionally, members of the OJI scientific committee examined the translations for technical issues related to the content.

Finally, we conducted a comparison between human translations provided by members of the OJI scientific committee and those generated by LLMs, to evaluate how closely the LLM translations align with expert human translations.

2.1. LLM Selection

For our experiments, we selected multiple state-of-the-art LLMs that are widely used for natural language processing and problem-solving tasks:

- **Llama 3.1 8B**: A lightweight LLM optimized for efficiency and performance (Llama Team, 2024).
- **Llama 3.2 3B**: A really lightweight LLM optimized for efficiency and performance (Llama Team, 2024).
- **GPT-4o**: The versatile flagship from Azure OpenAI, highly capable, known for its broad generalization abilities (OpenAI et al., 2024).
- **OpenLLMRo**: We tried multiple models from the OpenLLMRo community (Masala et al., 2024).
- **panSophic-1 preview**: A Romanian language-specific model.
- **Aya35B**: Aya 35B, released by Cohere, is part of a new family of state-of-the-art multilingual models. One of the 23 languages that it supports is Romanian (Aryabumi et al., 2024).

We conducted additional tests on Mistral 7B, Gemma 7B, and Gemini 1.5Pro, but the results for these models were below average.

Most of our models fit into memory and no quantization was needed except for the Aya35B model. According to (Marchisio et al., 2023), Aya35B shows minimal degradation in translation tasks (only -0.7% on the Flores benchmark).

GPT-4o was chosen for being state of the art, while the open-weight models were selected for their perceived strong performance in translating or knowledge of Romanian.

2.2. Translation Process

We translated each task using the LLMs listed in the previous section.

Peeperkorn et al. (2024) investigate the impact of temperature adjustments on the creativity of outputs produced by large language models. The study demonstrates that setting the temperature above 1 has minimal effect on the results, specifically in the context of story generation.

For each task and model, we initially performed one translation at a temperature setting of 0.6. For the top-performing models, we repeated the translations at varying temperature settings: 0.2, 0.6, and 1.0 to assess performance across different levels of randomness and across multiple iterations.

We started with a temperature of 0.6 as a balanced approach, providing a mix of randomness and certainty in the model's predictions. This allowed us to get a general sense of the model's performance. The top-performing models were then tested at different temperature settings to further understand their capabilities. A temperature of 0.2 was chosen to see how the model performs when it is more confident in its predictions. This could potentially lead to more accurate translations, but it could also result in less creative or diverse outputs.

On the other hand, a temperature of 1.0 was chosen to push the model towards more randomness in its predictions. This could lead to more diverse and creative translations, but it could also result in less accuracy.

2.2.1. Prompt

The following simple prompt was used for translation with all models:

You will be provided with an OJI (Olimpiada Județeană de Informatică) challenge and you will need to translate it from Romanian to English.

Rules: Any strings related to the body of the challenge should not be translated, such as input/output examples or file names.

The translation should follow the original format (numbering of paragraphs, examples, etc). The translation should be in the same tense as the original text. Respond only with the challenge body in English.

This prompt was chosen for its specific requirements, including the need to preserve the structure, strings, and tense of the original text. These constraints make the task an interesting subject for studying translation strategies and the challenges of translating technical and domain-specific content.

We aimed for a simple, reasonable prompt that typical LLM users could employ. However, as shown later, a more complex prompt may yield better results.

Moreover, the task’s focus on the OJI challenges also provides an opportunity to explore the translation of educational and competitive programming content, which is a less-studied area in translation studies.

2.3. Problem Selection

Most of the tests have been done on problems of low to medium difficulty, as our main method of automatic verification was to have the translated tasks solved by LLMs. We focus our analysis on 8th grade problems, which can be found on Kilonova (Kilonova problem lists). Kilonova is a renowned Romanian online judge platform, specifically designed for training in computer science olympiads. It has gained significant popularity in recent years due to its comprehensive collection of problems from all past olympiads. The platform provides an interactive and challenging environment for students to enhance their problem- solving and programming skills.

8th grade problems tend to have complex texts with many subproblems. The 8th grade curriculum also includes string-related problems, and we know that such problems tend to cause difficulties in translation due to commonly mistranslated terms, such as: *subșir* (subsequence, not necessarily a contiguous part of the sequence) and *subsecvent, ă* (subarray or substring).

We offer translations for all problems in the OJI dataset (Dumitran et al., 2024) and provide an external link to a broader translation dataset.

3. Translation Comparison

We start by emphasizing that, to our surprise, translating from Romanian to English does not seem to cause performance issues. Moreover, there may even be benefits to translating the task into English.

Figure 1 illustrates the difference between the maximum translation score across all models and the `ro_score`, showing this difference for each problem individually. For 22 out of 38 problems, the `ro_score` matches the maximum translation score, while for the remaining cases, the maximum translation score exceeds the initial score.

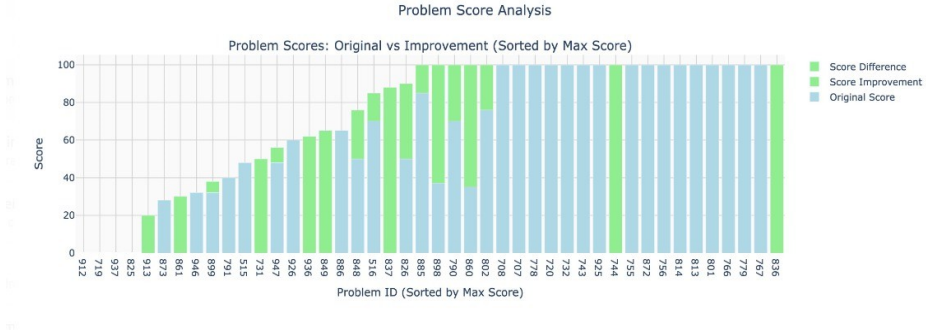


Figure 1: Comparison of maximum translation scores versus the original score (ro_score) across all problems

3.1. Automated Verification

Our methodology involved conducting five runs using GPT-4o to solve the initial Romanian tasks. After obtaining the baseline scores, the problem statements were translated into English using various LLMs and temperatures. Although increasing the number of runs improves the likelihood that an LLM will solve a given problem, it also incurs additional costs. Based on our findings in (Dumitran et al., 2024), we conclude that five runs provide a good balance between quality and cost. For lower temperatures, fewer runs are also effective.

The translated versions were then fed back into GPT-4o, where five additional runs were performed for each translated text. We compared the results of these runs with the initial Romanian versions. If a similar or better score was achieved on the translated English version, we considered the translation to be effective. In some cases, better scores were observed for the English translations due to several factors, including:

- **Randomness in the Outputs:** We often used temperatures higher than 0, introducing randomness that can sometimes lead to better outcomes in the English version.
- **Superior Training Data in English:** The model is typically trained on larger and more diverse English datasets, leading to better performance in English tasks.
- **Refined Error Handling in English:** LLMs generally have more robust error detection and correction mechanisms when processing English inputs, improving their overall problem-solving capabilities.
- **Better Understanding of Formal Structures in English:** English-specific models are better at understanding and interpreting the formal language structures often used in technical problems.

3.1.1. Overall Results

Figure 2 presents a comparative analysis of the 15 highest-scoring model runs, arranged in descending order of performance. The x-axis identifies each model run with an alphanumeric code, which represents specific model architectures, temperature settings ('t'), iteration number ('i'), and, where known, the number of parameters. The percentages above the bars show the performance difference compared to the Ro_score.

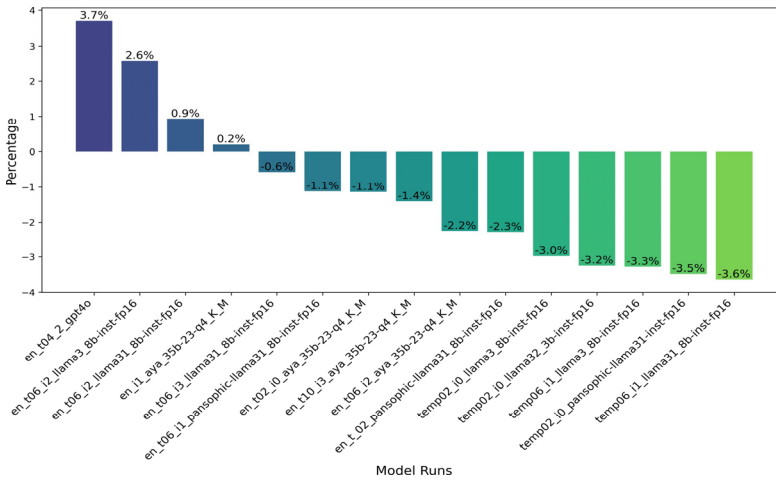


Figure 2: Top 15 model runs vs RoScore

The close clustering of scores, ranging from +3.7% to -3.6%, illustrates that top models achieved similar performance levels, with only marginal variations between them. Unsurprisingly, multiple runs with the same settings can yield different results due to the inherent randomness in model outputs. Open-LLmRo models were excluded from this analysis as their focus on Romanian significantly impacted their English translation capabilities.

These results underscore the diminishing returns at the upper end of model performance and suggest that further progress in enhancing translation accuracy may depend on fine-tuning models for specific tasks or further improving prompt design, rather than purely increasing model size or complexity.

3.2. Temperatures – the Creativity Parameter

Table 1: Average result on dataset over temperature and size

Translation Model	Temperature	Size	Scores	Score Average
gpt4o	0.6	*	58.84	58.84
ro_score	0.6	*	55.13	55.13

Translation Model	Temperature	Size	Scores	Score Average
aya	0.2, 0.6, 1.0	35B	54.00, 55.34 , 53.73	54.35
llama3	0.2, 0.6, 1.0	8B	52.16, 57.73 , 47.45	52.44
llama31	0.2, 0.6, 1.0	8B	50.05, 56.07 , 49.77	51.96
llama32	0.2, 0.6, 1.0	3B	51.89 , 51.41, 49.41	50.90
pans-llama31	0.2, 0.6, 1.0	8B	52.86, 54.02 , 46.93	51.27
gemini	0.6	*	45.11	45.11
mistral	0.6	7B	44.50	44.50
gemma	0.6	7B	43.77	43.77

For the models for which we ran more iterations, the scores are aligned respectively to the temperature in Table 1. We chose temperatures of 0.2, 0.6, and 1.0, as discussed in 2.2. We only ran temperature 0.2 and 1.0 for the LLMs with good results on temperature 0.6. Pricier models such as gpt4o or gemini were run only once, with 0.6. Key findings:

1. **(Optimal Temperature):** On average temperature 0.6 showed best results followed by 0.2 and 1.0.
2. **(Small Model):** Llama 3.2 3B, our smallest model, was the only one that performed better at a temperature of 0.2, indicating that smaller models tend to deviate from optimal results more quickly.
3. **(Temperature 0.6 Similar Text).** We add two translations by Llama 3.1 at temperature 0.6:

A word consisting only of small letters is given. We call anagram a word formed by rearranging the letters of the given word. For example, armata is an anagram of tamara. Obviously, a word can be considered its own anagram.

A word consisting only of lowercase letters is given. We call anagram a word formed by the letters of the given word, rearranging them if necessary. For example, armata is an anagram of tamara. Obviously, a word can be considered an anagram of itself.

Note that while the story is a bit modified in translation the rest of the text (not included here) is almost identical (the part related to the actual task and the restrictions). Thus, models tend to produce similar but slightly varied translations at temperature 0.6.

3.3. Model Size

We utilized language models of varying sizes to represent different capacities and computational requirements (see 1). The models, ordered by size, include Llama 3.2 3B, Gemma 7B, Mistral 7B, Aya 8B, Llama 3 8B, Llama 3.1 8B, Pansophic (based on Llama 3.1 8B), Aya 35B, and GPT-4o (estimated at over 100 billion

parameters). By incorporating models ranging from 3 billion to over 100 billion parameters, we investigated how model size impacts translation performance across temperature settings. This range allowed us to examine the trade-offs between computational resources and translation quality, as well as how different-sized models respond to temperature adjustments in language translation tasks.

Key Observations

1. GPT-4o at 100B+:

- GPT-4o, with over 100 billion parameters, demonstrates the highest performance in our study. However, GPT-4o's scores should be evaluated in comparison with other large models like Llama 3.1 405B. Additionally, GPT-4o's performance may be influenced by the fact that it was both the translation model and the judge, potentially giving it an advantage by aligning the translation with the evaluation model's own language patterns.

2. Smaller Models Struggle with Longer Tasks

- In our analysis, smaller models often struggled to complete translations, especially for longer, more complex tasks. Many of the tasks that received a score of 0 points after translation were incomplete. This was particularly evident in problems with a lot of restrictions and examples like problem 515, where examples and restrictions were either omitted or poorly translated. For instance, in one Aya translation, a big part of the text was omitted and instead the following text offered: The rest of the challenge remains unchanged and can be found in your original text, instead of translating the task and restrictions fully.

3. Llama Models' Performance at 8B and 3B

- The Llama models, known for robust multilingual capabilities, consistently succeeded in translation tasks, often matching or exceeding the original Romanian scores (Ro_score). Zero scores were only due to incomplete translations. This suggests that with better prompts tailored to smaller models and allowing multiple attempts, both Llama 3 and Llama 3.1 could potentially match GPT-4o's performance.
- However, Llama 3.2 3B underperformed in comparison to the other Llama models, proving inadequate for this specific task.

4. Similar Translation Quality Across Top Models:

- As detailed in Section 4, our qualitative analysis revealed no significant differences in translation quality among the top-performing models. The variations in performance were largely attributed to model fatigue when handling longer tasks.

4. Grammatical Analysis

Various translation techniques, including expansion, adaptation, transposition, and structure shift, continue to be employed by machine translation (MT) systems. As noted by (Wu et al., 2016), these systems utilize large datasets and deep learning algorithms to effectively implement these techniques, aiming to deliver high-quality translations across a wide range of languages and contexts, ultimately improving both the accuracy and readability of the output. Despite its widespread use and significant advancements, as Karami (2014) points out for Google Translate, the LLMs continue to encounter difficulties with more complex texts, idiomatic expressions, and languages that have less digital representation. These limitations underscore the necessity for ongoing improvements and suggest caution when depending on machine translation for critical or nuanced content.

The initial step in addressing the errors identified in the translation of algorithmic and informatics problems from Romanian into English by large language models (LLMs) is to classify them systematically. Classification enables the organization of errors into distinct categories, facilitating a more focused analysis of the translation issues. By grouping errors – such as mistranslations, inconsistencies in technical terminology, or the omission of critical computational terms – it becomes possible to identify recurring patterns and pinpoint areas where LLMs struggle the most. This structured approach lays the foundation for subsequent error quantification and qualitative analysis, essential for improving the over- all performance of automated translation systems in technical domains. (Lommel et al., 2014) introduce a hierarchical translation error taxonomy known as MQM, which distinguishes between fluency errors and accuracy errors. According to Lommel et al. (2014), fluency errors are defined as issues “related to the language of the translation, irrespective of its status as a translation,” while accuracy errors pertain to “how accurately the target text reflects the content of the source text.” Using these categories, errors can be classified as either minor or critical. For this study, errors were categorized based on the classifications established by researchers such as Hemchua and Schmitt (2007), Hsu (2014), Costa et al. (2015), and according to the frequency of errors identified in the data.

The following table provides a comparative analysis of translation errors identified across various large language models (LLMs) such as Mistral, Llama, Gemma, Gemini, GPT-4o, and Aya. Each model was tasked with translating technical structures from Romanian into English, with the table highlighting the errors, the incorrect target language structure, and the suggested correct translations. While the table does not include every error identified, it highlights those that significantly impact the quality of the translation.

In this study, several algorithmic problems were translated from Romanian into English by various large language models (LLMs), including Mistral, Llama 3, Llama 3.1, Gemma, Gemini, GPT-4o, and Aya. A consistent pattern of errors and systematic distortions was observed across these translations.

Table 2: Comparative analysis of translation errors across LLM models

LLM Model	Problem Number	Structure in Source Language	Error Type
Mistral	515	rectiliniu	Lexical error
Mistral	802	număr de ordine	Semantic error
Llama	298	o matrice pătratică	Omission error
Mistral, Llama, Gemma, Gemini, GPT-4o	All texts	„x”	Punctuation error
Mistral, Llama, Gemma, Gemini, GPT-4o	All texts	Tables and their contents	Structural error
GPT-4o	899	șir	Semantic error
Aya	299	curent	False friend
GPT-4o	802	pozițiile de început și de final ale acestor secvențe în șirul din set	Semantic error

Notably, certain parts of the text were left untranslated. For instance, the section “restricții și clarificări” was often only partially translated as “restrictions,” omitting “clarifications.” Additionally, some tasks were either partially translated or left untranslated entirely, as demonstrated by Llama 3’s translation of problem 899. Furthermore, all LLMs failed to fully translate the content of tables, with at least one column from the “example” sections consistently omitted, and the explanatory content entirely absent. In addition to these omissions, punctuation errors were also frequent, with quotation marks, mathematical symbols, and other punctuation marks being incorrectly rendered. These recurring issues highlight the limitations of LLMs in accurately translating structured, technical content.

Several types of translation errors were identified in the output of LLMs when translating algorithmic problems from Romanian into English. One prominent error occurred in the translation of the term “rectiliniu” in problem 515, which was consistently rendered as “straight” by all the LLMs, instead of the correct term “linear”. This constitutes a lexical error, as the incorrect word choice alters the intended meaning within the context of the problem. Additionally, this could be classified as a semantic error, given that the LLMs failed to capture the appropriate technical meaning of “rectiliniu” in the domain of algorithmic and computational language, where “linear” is the precise term. The misinterpretation likely arises from the general meaning of “rectiliniu” in everyday language, where “straight” is a common equivalent, but it is not suitable in the technical context, leading to a distortion of the problem’s meaning. In the study, Llama exhibited a notable translation error in problem 298 by omitting the adjective “pătratică” when translating the Romanian structure “o matrice pătratică” as “a matrix”, rather than the correct “a square matrix”. This constitutes a lexical omission error,

where a critical modifier – in this case, the adjective “square” – was not translated, resulting in an incomplete and less precise rendering of the source text. The omission impacts the technical accuracy of the translation, as “square matrix” is a specific term in mathematics that conveys essential information about the matrix’s dimensions. The failure to include this detail diminishes the clarity and correctness of the translated problem, potentially leading to misinterpretation of the task.

Discrepancies were observed in the translation of key technical terms, particularly the Romanian word “șir”, which was inconsistently rendered by different LLMs. For instance, in problem number 899, Llama and GPT-4o translated “șir” as “sequence”, while Llama 3 translated it as “string”. While both translations capture certain aspects of the term, neither fully aligns with the technical context in which “șir” is used. The most accurate translation in this context would be “array”, which is a more precise term in algorithmic and computational problems. According to the literature (Hopcroft et al., 2006), “string” refers to a sequence of characters that can potentially have a very long or infinite length, which may not align with all the contexts given in the algorithmic. The use of “sequence”, while closer, still lacks the specificity required to convey the array-like properties of “șir”. Thus, the preferred translation of “șir” in such cases would be “array”, as it more accurately reflects the intended data structure within the problem’s context.

In conclusion, the large language models (LLMs) in this study, especially Llama3.1, Llama 3.2, GPT- 4o, and Aya23 – showed strong proficiency in translating algorithmic problems from Romanian to English, with overall translation quality approaching that of human translators. However, recurring issues like lexical mismatches, semantic inaccuracies, omissions, and structural inconsistencies highlight the need for human revision to ensure accuracy in technical domains. While LLMs are valuable for initial translations, human oversight remains crucial, especially in specialized contexts like algorithmic problem translation.

5. Working Solution

We enhanced the OJI Romanian dataset by incorporating manually verified English translations, creating a valuable benchmark for future research in multilingual LLMs. Drawing on extensive experience with Romanian competitive programming, we identified and corrected common machine translation errors, particularly technical jargon often mistranslated by generic tools. Key terms such as “cerință” (task), “subsecvență” (subarray), “subșir” (subsequence), and “șir de caractere” (string) were carefully handled to ensure accuracy.

Our experiments primarily used GPT-4o and the OpenAI API. We began with a basic prompt – “Can you translate this problem statement into English in the context of competitive programming?” – which, although imperfect, significantly improved focus on competitive programming terminology compared to generic translation approaches.

We further improved the translations by manually correcting recurring errors and refining the prompt based on our expertise with the OJI dataset. This refined prompt ensured a consistent structure in translated problem sets and accommodated Markdown and LaTeX formatting, enhancing clarity and functionality.

For example, a simple prompt that doesn’t account for specific terms can lead to pitfalls. Misuse of terms like ‘subsequence’ and ‘substring’, or inconsistent descriptions of printed data, can cause confusion. In competitive programming, even one misunderstood word can result in solving an entirely different problem.

Felicia este interesată de subșirul maxim lexicografic al unui șir de caractere. Rețineți că un șir a este considerat mai mic în ordine lexicografică decât un șir b dacă a este prefix al lui b , sau dacă există o poziție i pentru care avem $a[1] = b[1], \dots, a[i-1] = b[i-1]$, și $a[i] < b[i]$. Astfel, subșirul maxim lexicografic al unui șir de caractere este cel mai mare subșir, în ordinea lexicografică, al unui șir de caractere (de exemplu $zzxx$ pentru $azbxazbxaax$). Pentru un șir s de caractere vom nota cu $m(s)$ subșirul maxim lexicografic al lui s , și cu $v(s) = |m(s)|$ lungimea acestui subșir.

Felicia vă dă un șir s format din caractere mici ale alfabetului englez. Considerați toate subsecvențele continue s' ale lui s . Felicia vrea să calculați suma valorilor $v(s')$ pentru toate subsecvențele posibile s' amintite anterior.

Figure 3: Original Kilonova problem 15

Requirement

Felicia is interested in the maximum lexicographic substring of a string of characters. Note that a string a is considered to be less than a string b in lexicographic order if a is a prefix of b , or if there exists a position i such that $a[1] = b[1], \dots, a[i-1] = b[i-1]$, and $a[i] < b[i]$. Thus, the maximum lexicographic substring of a string of characters is the largest substring, in lexicographic order, of a string of characters (for example, “zzxx” for “azbxazbxaax”). For a string s of characters, we will denote the maximum lexicographic substring of s as $m(s)$, and the length of this substring as $v(s) = |m(s)|$.

Felicia gives you a string s consisting of lowercase letters of the English alphabet. Consider all continuous subsequences s' of s . Felicia wants you to calculate the sum of the values $v(s')$ for all possible subsequences s' mentioned above.

Figure 4: Translated Problem Statement with issues

5.1. Efficient Prompt

The final version of the prompt used is: “Please process the following text according to the specified instructions: You will be given a competitive programming problem statement in markdown, written in the Romanian language, using GFM extensions and MathJax/LaTeX math between dollar signs (\$ or \$).

Another extension is the fact that image attachments are defined using a syntax similar to [name.png], with optional attributes named after the end with a vertical bar ('|'). You must translate the statement in the English language, while preserving mathematical values, variable names, general syntax, structure and format. You must also preserve the custom image format exactly as is. The word Cerint, a~ is always translated to Task, Date de intrare is translated to Input data, Date de ieşire is translated to Output data, subsecvență is translated to subarray, subșir is translated to subsequence, Restricții și precizări to Constraints and clarifications, vector is translated to array, șir de caractere is translated to string. In addition, in the Date de intrare and Date de ieşire sections, if you see expressions such as Pe prima linie, Pe a doua linie etc., you want to use the verb contain to describe the data we need to read or print. In the Date de ieşire sections, you can also use the verb print to describe the data we need to print. When separating large integers (especially in latex/mathjax math) in groups of 3 digits, do not add a comma. Instead, add a backslash followed and preceded by a single space character. After you are done with translating, please double check the statement and fix potential grammar and/or syntax errors according to the rules of English language.”

Using the refined prompt, the translations became nearly flawless, with no significant errors detected by linguists or members of the competitive programming scientific committee. The translations met both linguistic and technical standards.

The final version of the OJI problem below illustrates the improvements achieved through prompt engineering and domain-specific adjustments while preserving competitive programming terminology.

Felicia is interested in the maximum lexicographic subsequence of a string. Note that a string a is considered smaller in lexicographic order than a string b if a is a prefix of b , or if there exists a position i for which we have $a[1] = b[1], \dots, a[i-1] = b[i-1]$, and $a[i] < b[i]$. Thus, the maximum lexicographic subsequence of a string of characters is the largest subsequence, in lexicographic order, of a string of characters (for example `zzxx` for `azbxaazbxaax`). For a string s of characters, we denote by $m(s)$ the maximum lexicographic subsequence of s , and with $v(s) = |m(s)|$ the length of this subsequence.

Felicia gives you a string s consisting of lowercase English letters. Consider all continuous subarrays s' of s . Felicia wants you to calculate the sum of the values $v(s')$ for all possible subarrays s' mentioned earlier.

Figure 5: Correct Translation using enhanced prompt

Given the relatively large size of the dataset—over 300 problems, as shown by this list from the Kilonova online judge—relying solely on manual browsing was not feasible.

Automating the translation process became essential. We obtained markdown versions of the problem statements from Kilonova, the only Romanian online judge hosting the entire OJI dataset. These files enabled us to develop Python scripts to process the raw statements, interact with the OpenAI API, and improve

formatting, reducing potential errors related to the LLM’s handling of markdown syntax.

Once the dataset was fully translated, we uploaded it to a GitHub repository for evaluating model performance. We further improved the translations by fixing markdown issues and refining specific aspects. This resulted in a comprehensive collection of Romanian problems accurately translated into English, providing a valuable resource for future research and evaluation.

6. Conclusions

This study demonstrates that large language models (LLMs) can effectively translate complex technical content from Romanian to English, achieving quality comparable to human translators when provided with well-crafted prompts and human oversight. While focused on Romanian to English, our findings open doors for exploration in other language pairs. Automated translation in STEM competitions is a particularly promising application, where translation quality is crucial and manual translations have historically faced issues of accuracy and potential fraud, highlighting the need for reliable automated solutions.

The results indicate that models like GPT-4.0 and Llama 3.1 8B consistently perform well, though smaller models struggle with complex tasks. Proper prompts and temperature settings (such as 0.6) proved essential for optimal performance. However, despite advancements, lexical errors and structural inconsistencies – especially in smaller models – suggest that human oversight remains necessary.

By enhancing the OJI dataset with manually verified translations, this study provides valuable resources for future research into multilingual LLMs. Expanding these efforts to other language pairs and refining the use of LLMs in high-stakes environments like educational competitions can significantly improve accessibility and fairness in such events.

References

- Abhimanyu Dubey, Abhinav Jauhri et al 2024. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783. <https://arxiv.org/abs/2407.21783>.
- Adrian Cosma, Ana-Maria Bucur, and Emilian Radoi. 2024. RoMath: A Mathematical Reasoning Benchmark in Romanian. arXiv preprint arXiv:2409.11074. <https://arxiv.org/abs/2409.11074>.
- Adrian Cosma, Bogdan Iordache, and Paolo Rosso. 2024. RoCode: A Dataset for Measuring Code Intelligence from Problem Definitions in Romanian. arXiv preprint arXiv:2402.13222. <https://arxiv.org/abs/2402.13222>.
- Adrian Marius Dumitran, Adrian Catalin Badea, and Stefan-Gabriel Muscalu. 2024. Evaluating the Performance of Large Language Models in Competitive Programming: A Multi-Year, Multi-Grade Analysis. arXiv preprint arXiv:2409.09054.

- Alexandre Costa, Ling Wang, Luís Tiago, Rui Costa, and Luisa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29:127–161.
- Arle Lommel, Aljoscha Burchardt, and Hans Uszkoreit. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'14)*, European Language Resources Association (ELRA), 14–20.
- J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffman, F. Song, and G. Irving. 2021. Scaling Language Models: Methods, Analysis and Insights from Training Gopher. arXiv preprint arXiv:2112.11446. <https://arxiv.org/abs/2112.11446>.
- J.-A. Hsu. 2014. Error classification of machine translation: A corpus-based study on Chinese-English patent translation. *Studies of Translation and Interpretation*, 18:121–136.
- John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. 2006. *Automata Theory, Languages, and Computation*. Addison-Wesley.
- Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2023. How Does Quantization Affect Multilingual LLMs?. arXiv preprint, arXiv:2407.03211. URL: <https://arxiv.org/abs/2407.03211>.
- Kilonova problem lists. Retrieved from https://kilonova.ro/problem_lists/456
- Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is Temperature the Creativity Parameter of Large Language Models?. arXiv preprint arXiv:2405.00492. <https://arxiv.org/abs/2405.00492>.
- Mihai Masala, Denis C. Ilie-Ablachim, Dragos Corlatescu, Miruna Zavelca, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. OpenLLM-Ro – Technical Report on Open-source Romanian LLMs. arXiv preprint, arXiv:2405.07703. URL: <https://arxiv.org/abs/2405.07703>.
- Omid Karami. 2014. The brief view on Google Translation Machine. *Seminar in Artificial Intelligence on Natural Language*, Germany.
- OpenAI, Josh Achiam, Steven Adler, et al. 2024. GPT-4 Technical Report. arXiv preprint, arXiv:2303.08774. URL: <https://doi.org/10.48550/arXiv.2303.08774>.
- Saowalak Hemchua and Norbert Schmitt. 2007. An analysis of lexical errors in the English compositions of Thai learners. *Linguistics*, Corpus ID: 42846941.
- Viraat Aryabumi, John Dang, Dwarak Talupuru, et al. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. arXiv preprint, arXiv:2405.15032. URL: <https://arxiv.org/abs/2405.15032>.
- Yonghui Wu, Mike Schuster, Zhifeng Chen et al 2016. Google’s neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. URL: <https://arxiv.org/abs/1609.08144>.

DETECTING FAKE NEWS USING BERT AND ROBERTA MODELS: A RESEARCH STUDY

ASMII SHRIPAD

*Cambridge Centre for International Research (CCIR), 184 Cambridge Science
Park Milton Rd, Milton, Cambridge CB4 0PZ, United Kingdom
asmiishripad@gmail.com*

Abstract

With the increase in the use of social media, the dissemination of fake news has become a serious problem. It has been reported in a previous study on the online news on Twitter that the spread of false news is six times faster than truthful content and that 70% of the users could not distinguish real from fake news. This study aims to combat the spread of fake news with the help of a fake news recognition model built using pre-trained models BERT and RoBERTa for fake news classification. The model also makes use of the hugging face transformers library as a baseline. The objective of this model is to assist users in identifying the reliability and validity of the news source by recognising the source of the news. The average accuracy of BERT is 95.95%, while RoBERTa has an accuracy of about 96.38%. The precision for the BERT and RoBERTa models is averaged around 95.50% and 96.45% respectively, while the recall is around 95.12% for BERT and 95.23% for RoBERTa. The models give fairly accurate predictions when tested on an alien dataset. The research paper dives deeper into the implementation, training and experiments conducted along with model evaluation.

1. Introduction

With the spike in social media usage over the years, there has also been a drastic increase in the instances of the spread of fake news. Due to this, many people tend to find themselves trapped in the web of lies and misinformation on online platforms. The dissemination of false news has drastic and long-lasting impacts on governments, individuals and organizations. This is now a major concern among various groups, owing to an exponential spike in social media usage since the COVID-19 pandemic. Therefore, it is necessary that the people are able to guard themselves against the spread of misinformation, while maintaining their social media presence. Identifying misleading news proves to be a challenging and cumbersome task.

This research paper outlines the usage of binary classification techniques included in Natural Language Processing (NLP) along with the pre-trained models

of BERT and RoBERTa to accurately differentiate between fake and real news snippets. The model mainly uses the title of the article to determine whether it is fake or real. An example of the title included in the dataset is: ‘Democrats Have a 60 Percent Chance to Retake the Senate - The New York Times’. The first impression after reading this title is that it corresponds to a real news article. However, in reality this is a fake news snippet. Therefore, fake news detection, though crucial and important in today’s times, can prove to be extremely complicated, nuanced and difficult to implement. The primary purpose of this research is to make predictions about news snippets in English. Further research and experimentation can be carried out on data that contains various other languages.

2. Background

There is a long history of work done in the field of Natural Language Processing for binary classification tasks. In the recent years, many new methodologies have come up for implementing the previously known NLP tasks.

The first research paper recognizes the need for fake news classification as it is becoming a big societal threat with its fast dissemination over the Web and social media along with its power to shape public opinion. The paper highlights that on a small amount of data, news title stylistic and linguistic features are better than the same type of features extracted from the news body in predicting fake news (Shrestha and Spezzano, 2021).

The introduction of the BERT model led to major changes in the implementation of text classification tasks including fake news detection. BERT stands for ‘Bidirectional Encoder Representations from Transformers’. It is a conceptually simple yet an empirically powerful model. BERT is designed to pre-train deep bidirectional representations from unlabelled texts by learning both left and right context in all layers. As a result, this pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks. (Jacob Devlin and Toutanova, 2019)

In a few other research papers, the fake news detection task is implemented as a Natural Language Interference task rather than the traditional binary classification approach to the problem. The model is individually trained on a number of the strongest Natural Language Interference (NLI) models such as Dense RNN, Dense CNN, and Gated CNN along with BERT (Kai-Chou Yang and Kao, 2019).

Exploring the correlations of publisher bias, news stance, and relevant user engagements simultaneously, provides us with valuable insights and encompasses a Tri-Relationship Fake News detection framework (TriFN), which is a tri-relation framework between the publisher, news and social engagements of the news article. This identified tri-relationship can also be used to classify fake news (Kai Shu and Liu, 2017).

3. Methodology

3.1. Dataset

The original dataset (SINHA, 2020) consists of a total of 20,800 rows, which was reduced after data cleaning and data refining techniques. The redundant values along with the null values were removed from the dataset reducing the count of the rows to 18285 rows. The dataset included features such as: title of the article, the author, the text present in the article and the label. The author column was removed from the dataset as it did not significantly contribute to the task of recognizing fake news. The labels or the classes in the dataset include 0 and 1 which correspond to FAKE and REAL respectively. There are a total of 10361 fake news snippets and 7924 real news snippets in the dataset. In both the experiments, the model makes use of the “title” feature to make predictions on the class or label of the data. The original data was then split into train, test and validation dataset with a size of 70%, 24% and 6% respectively. The distribution of ‘Fake’ and ‘Real’ labels selected for training, testing, and validation is slightly random. This has been done to avoid biasing the model training towards specific labels and to generate a generic model while training.

Table 1: Number of fake and real news snippets in the dataset

Label	Count
0 (Fake)	10361
1 (Real)	7924

The distribution of “Fake” and “Real” labels selected for train, test and validation are random.

Data preprocessing steps:

- Data Cleaning
- Tokenization
- Data Padding

Data Cleaning is a process where the existing data is refined and filtered out from its noise. It includes handling of missing values, smoothing of data and many other steps. This was done using the inbuilt methods in the pandas module in Python. Along with this, the tokenization and padding of data was done using the BERT module to match the input format required by the BERT and RoBERTa models. After these steps, the dataset was converted into the 'torch' format. This was done to ensure compatibility during training as it is implemented using the PyTorch library.

3.2. The BERT Model

BERT stands for ‘Bidirectional Encoder Representations from Transformers’. The BERT model is specifically classified as a Masked Learning Model that enables bidirectional learning. The BERT model is built upon the transformers’ architecture. The transformers model is based on the concept of ‘selective attention’ in human beings. It is used in supervised, unsupervised as well as deep learning tasks. Due to limited capacity of computing resources, the transformers architecture allows models to only focus and capture specific data relevant to the problem at hand. To perform this, transformers are made up of encoders and decoders. The data is first passed through the encoders that break down the data into several input features. These features then act as an input to the decoder. The decoder mainly selects the important input features and uses those to produce the output. Figure 1 illustrates the workflow of the BERT Model.

Since the BERT model is based on transformers, it can easily be fine-tuned using a particular dataset. With the help of techniques such as embedding, next sentence prediction and vectorization, it is used to perform NLP based tasks with considerable accuracy.

The BERT model workflow includes three major steps (see Figure 1).

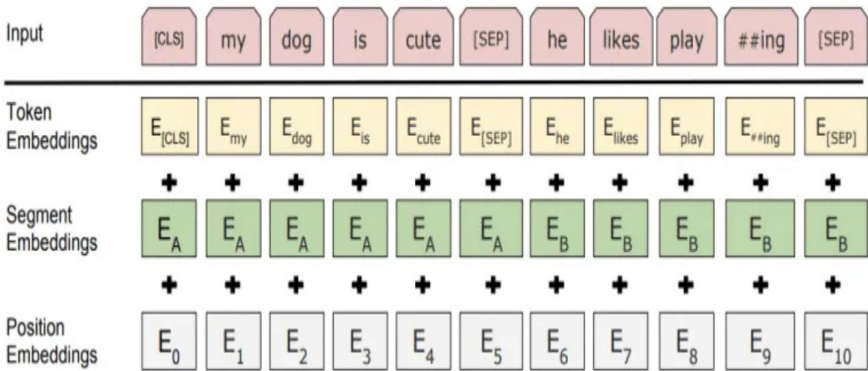


Figure 1: Workflow of the BERT Model (Jacob Devlin and Toutanova, 2019)

1. Token Embedding

Each word in the input text is converted into a number format with the help of the BERTTokenizer so that each word corresponds to a particular value.

2. Segment Embedding

It helps the model to understand the context, relationship and boundaries between 2 different segments of the sentence aiding in comprehension and understanding of the text.

3. Position Embedding

Position embedding helps the model to understand the position of different words in the sentence, enabling a better understanding of the grammar and structure of the sentence.

Bert Input Representation (Jacob Devlin and Toutanova, 2019):

$$\mathbf{H}_i^{(0)} = \mathbf{E}(x_i) + \mathbf{P}(i) + \mathbf{S}(i)$$

where $\mathbf{E}(x_i)$ is the token embedding, $\mathbf{P}(i)$ is the position embedding, x_i and $\mathbf{S}(i)$ is the segment embedding.

Transformers' Function (Ashish Vaswani and Parmar, 2023):

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

$$Q = W^Q \mathbf{H}_i^{(l)} \text{ (Query)}$$

$$K = W^K \mathbf{H}_i^{(l)} \text{ (Key)}$$

$$V = W^V \mathbf{H}_i^{(l)} \text{ (Value)}$$

where W^Q , W^K , W^V are projection matrices and d_k is the key/query vector

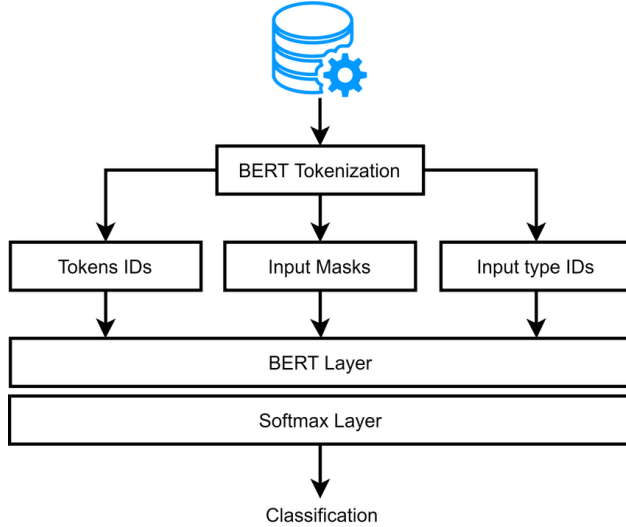


Figure 2: Fine tuning of the BERT Model (Rahali and Akhlouf, 2021)

3.3. The RoBERTa Model

RoBERTa stands for a ‘Robustly Optimized BERT Pre-training Approach’. It is similar to BERT and is also built on the self-attention process of the Transformers model. It optimizes the training and fine-tuning process used by BERT. RoBERTa makes changes in the key hyperparameters, the embeddings and the BERT model setup to provide enhanced results. The BERT feature of ‘next sentence prediction’ has also been removed in the implementation of RoBERTa. The RoBERTa model makes use of dynamic masking, large batch size, a byte-level vocabulary and sentence packing (sentences are packed together to reach 512 tokens) in contrast to BERT Architecture. It has been trained larger amount of data (160GB), as compared to BERT (13GB). Therefore, it has a richer vocabulary than BERT and is more likely to recognize tokens.

Thus, RoBERTa with its refined techniques for training and text classification, provides better and more accurate results in most scenarios.

3.4. Hyperparameter Tuning

3.4.1. Batch Size

The batch size specifies the number of samples that the model should work through before updating its internal model parameters. It is also observed that an increase in the training batch size and decay in learning rates have similar effects on the model performance. Setting a very high or low value of the batch size can also lead to the randomization of 5 results. A general trend noted across models is that an increase in batch size leads to a lower performance (Samuel et al., 2018). A higher batch size is known to significantly impact the model’s ability to generalize the results. This is mainly because the smaller batches update their internal parameters a greater number of times than compared to larger batches, ultimately leading to a lower ability to generalize. Thus, the experiments mentioned below make use of wide-ranging batch size values to find out the optimum value that yields the highest model performance.

3.4.2. Weight Decay

Weight decay, or also known as L2 regularization, is a technique employed to penalize the complexity of a model. Interactions between the various hyperparameters tend to increase the complexity of the model. Weight decay values make sure that the model does not become too complex by penalizing the high values of weights, during the calculation of the loss function. Weight decay is a very powerful and important hyper parameter that also helps to reduce over fitting, increase model stability and also improve the generalization in over parameterized models (Mudadla, 2023).

$$\text{Loss Function (after weight decay)} = L(w, b) + \frac{\lambda}{2} \|\mathbf{w}\|^2$$

where $L(w, b)$ is the loss function with the weight vector w and bias b before adding the weight decay, λ is the regularization coefficient and $\|\mathbf{w}\|^2$ is the sum of the square of weights.

4. Experiments and Analysis

The various experiments described below were conducted by changing the values of different hyper-parameters to see how they affect the accuracy and performance of the model. Overall, the results are categorized as training and validation, evaluation and random testing. At the end, unseen news snippets have been passed to both the models with the optimal hyper parameter values found from the above experiments. The results are summarized below.

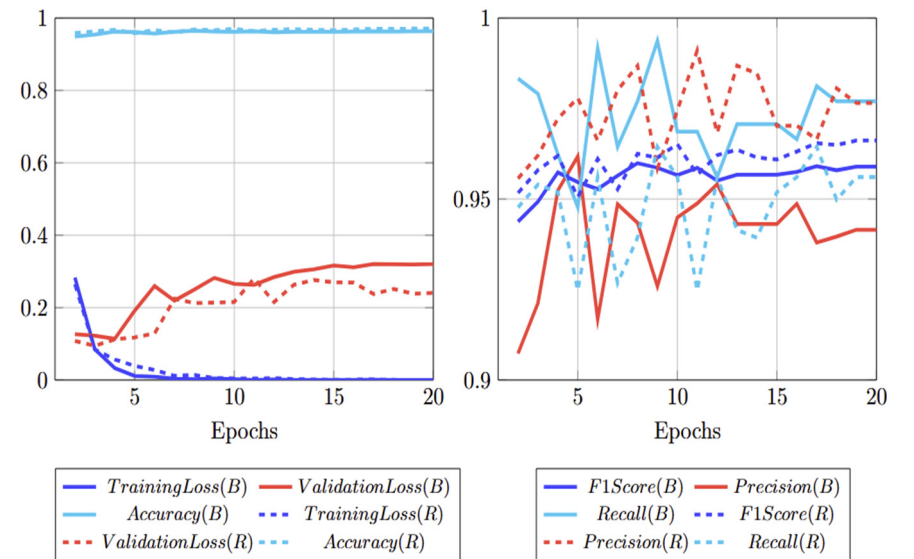
4.1. Training and Validation Results

There are three experiments conducted with varying training batch size of 4, 32 and 128. The warm up steps, number of epochs and weight decay are kept constant at 128, 20 and 0.01 respectively in the first set of experiments. In the next set of experiments, the batch size is kept constant at 32 along with the warmup steps and the number of epochs. The weight decay values are changed subsequently to 0.1 and 1.0 for the next two experiments. Figures 3 to 7 illustrate the model training and validation results for each of the experiments.

The Figures 3(a), 3(b), 3(c) and 3(d) display the testing and validation results for the BERT and RoBERTa models. The training loss for both models decreases steadily with each passing epoch. The F1 score, Recall and Precision for both the models is in the range of 0.9 and 1. The confusion matrices for both the models validate the results displayed in the above graph.

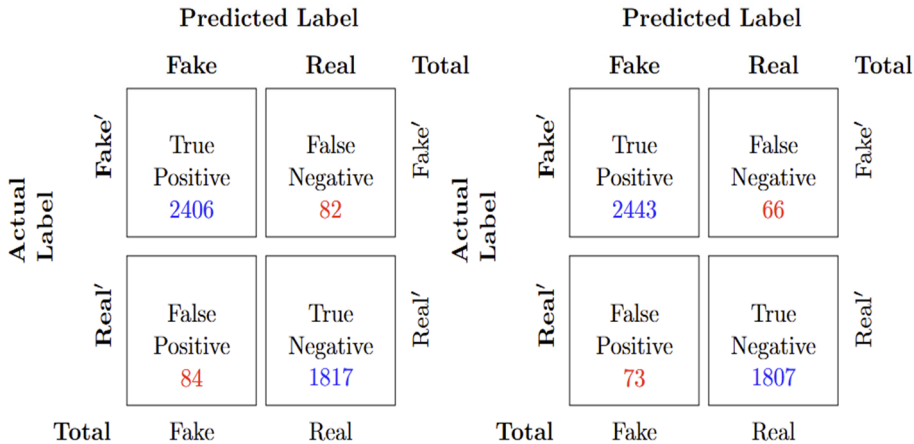
The Figures 4(a), 4(b), 4(c) and 4(d) display the performance of the BERT and RoBERTa models with a batch size of 32 and weight decay of 0.01. Figure 4(a) shows an exponential decrease in the training loss for both the models over subsequent epochs. Figure 4(b) highlights the F1 score, recall and precision for both the models. The confusion matrices underscore the above results.

Figures 5(a), 5(b), 5(c) and 5(d) show the results for the BERT and RoBERTa models with batch size of 128 and weight decay of 0.01. Figure (a) displays the steady decrease in the training loss of both the models over subsequent epochs. Figure (b) shows the fluctuations of the aforementioned values in the first few epochs which stabilize after the completion of 10 epochs. The confusion matrices in figures 5(c) and 5(d) highlight the above results.



(a) Training Results for BERT (B) and RoBERTa (R)

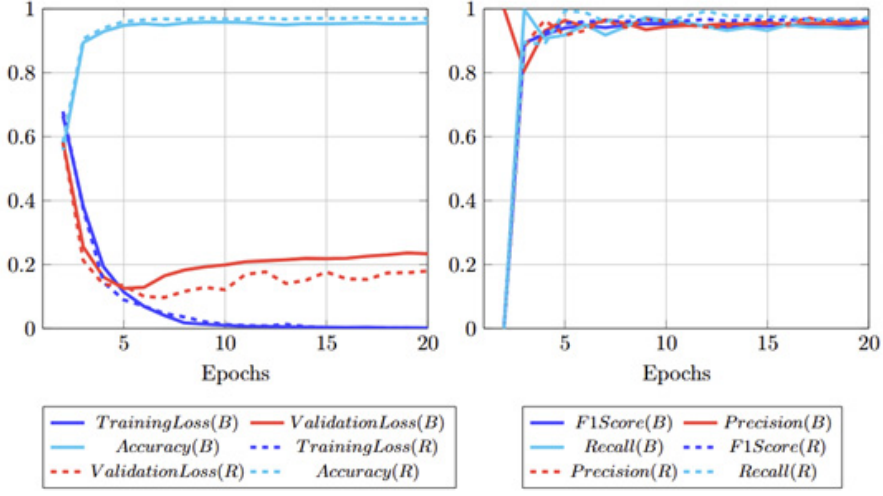
(b) Validation Results for BERT (B) and RoBERTa (R)



(c) Confusion Matrix (BERT)

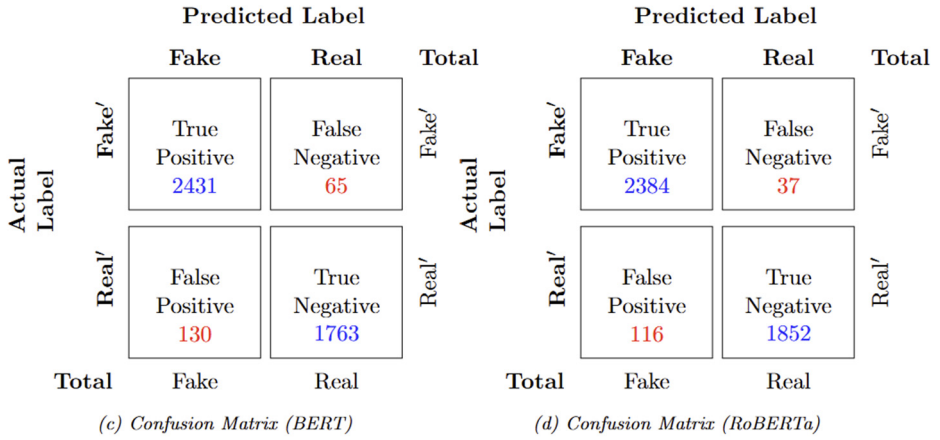
(d) Confusion Matrix (RoBERTa)

Figure 3: Result with training batch size=4, epochs=20, weight decay=0.01



(a) Training Results for BERT (B) and RoBERTa (R)

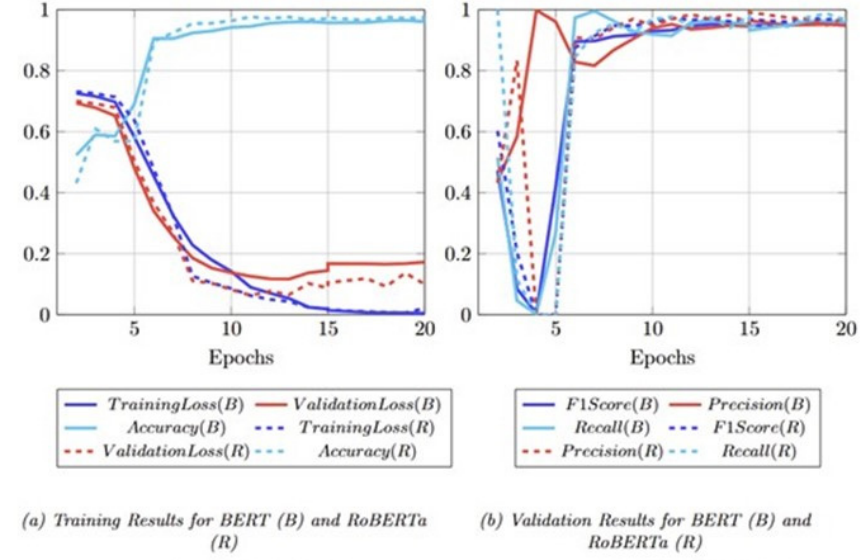
(b) Validation Results for BERT (B) and RoBERTa (R)



(c) Confusion Matrix (BERT)

(d) Confusion Matrix (RoBERTa)

Figure 4: Result with training batch size=32, epochs=20, weight decay=0.01



		Predicted Label		Total			Predicted Label		Total	
		Fake	Real				Fake	Real		
Actual Label	Fake'	True Positive 2364	False Negative 104	Fake'	Actual Label	Fake'	True Positive 2403	False Negative 76	Fake'	
	Real'	False Positive 102	True Negative 1819			Real'	False Positive 81	True Negative 1829		
Total		Fake	Real			Total	Fake	Real		

(c) Confusion Matrix (BERT)

		Predicted Label		Total			Predicted Label		Total	
		Fake	Real				Fake	Real		
Actual Label	Fake'	True Positive 2403	False Negative 76	Fake'	Actual Label	Fake'	True Positive 2403	False Negative 76	Fake'	
	Real'	False Positive 81	True Negative 1829			Real'	False Positive 81	True Negative 1829		
Total		Fake	Real			Total	Fake	Real		

(d) Confusion Matrix (RoBERTa)

Figure 5: Result with training batch size=128, epochs=20, weight decay=0.01

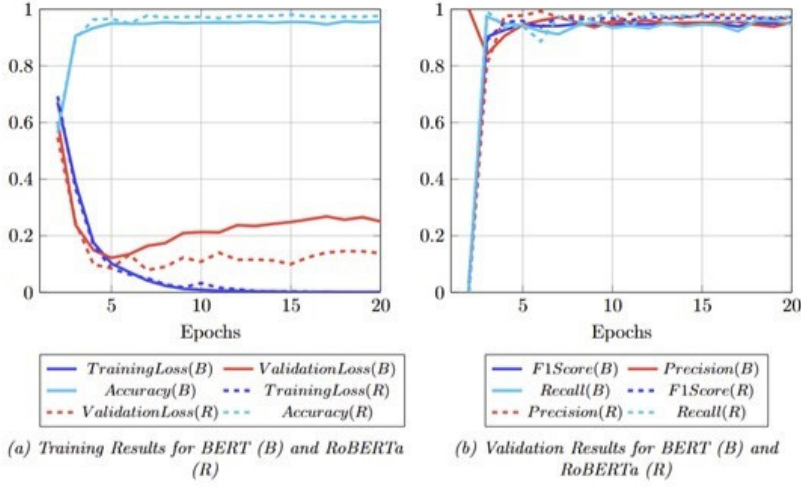


Figure 6: Result with training batch size=32, epochs=20, weight decay=0.1

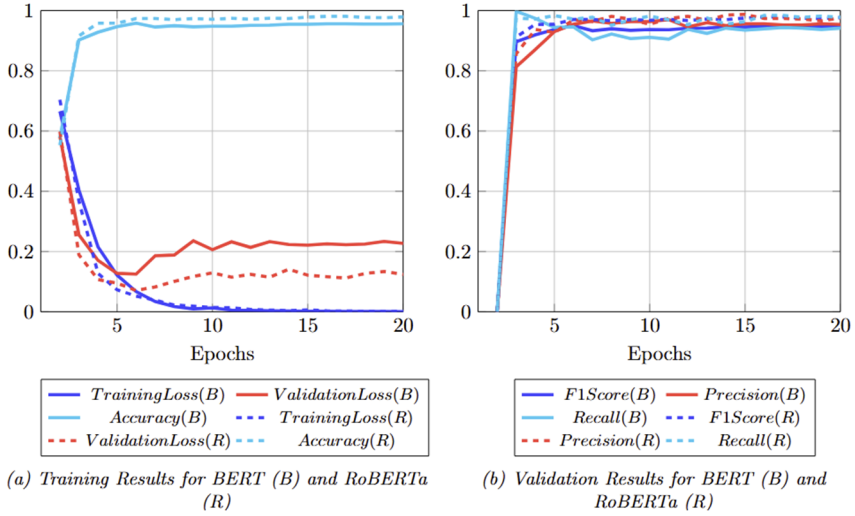


Figure 7: Result with training batch size=32, epochs=20, weight decay=1.0

Figures 6(a) and 6(b) demonstrate the results for the BERT and RoBERTa models when the batch size is 32 with a weight decay of 0.01. The training loss decreases exponentially with each epoch. Figure 6(b) shows the validation results for the BERT and RoBERTa models with a sharp rise in values for the first few epochs. Figures 7(a) and 7(b) consist of the results of the BERT and RoBERTa models with a batch size of 32 and weight decay of 1.0. These results too show a similar trend in values as in the previous experiment.

4.2. Test Data Evaluation Results

Tables 2 summarizes the test data evaluation statistics for the experiments with varying batch sizes whereas tables 4 encapsulates the results of the experiments with varying weight decay values for BERT and RoBERTa models. Table 4 demonstrates the average inference time taken by these models to process an input data. However, it must be noted that these results are only for news snippets in the English language. For other languages, another set of hyperparameters can be determined with further experimentation.

Table 2: Validation Results for BERT and RoBERTa with varying batch size

Metric	BERT			RoBERTa		
	Batch Size (4)	Batch Size (32)	Batch Size (128)	Batch Size (4)	Batch Size (32)	Batch Size (128)
Accuracy (%)	96.21	95.55	95.30	96.83	96.51	96.42
F1 Score (%)	95.63	94.75	94.60	96.23	96.03	95.88
Recall (%)	95.58	93.13	94.69	96.11	94.10	95.75
Precision (%)	95.68	96.44	94.59	96.47	98.04	96.01

Table 3: Validation Results for BERT and RoBERTa with varying weight decay

Metric	BERT			RoBERTa		
	Weight Decay (0.01)	Weight Decay (0.1)	Weight Decay (1)	Weight Decay (0.01)	Weight Decay (0.1)	Weight Decay (1)
Accuracy (%)	95.55	97.08	95.64	96.51	95.32	96.85
F1 Score (%)	94.75	96.53	94.8	96.03	94.64	96.33
Recall (%)	93.13	96.11	96.11	94.10	94.07	96.12
Precision (%)	96.44	96.95	93.87	98.04	95.21	96.53

Table 4: Average Inference Time for the Models

Model	Average Inference Time (milliseconds)
BERT	2.32
RoBERTa	3.72

4.3. Random Test Results

Table 5 summarizes the results when two random unseen new snippets are passed as input to the BERT and RoBERTa models. The news real news snippet is taken from the latest news published by 'The New York Times'. The confidence level for each prediction has been derived using the softmax activation function.

Table 5: Random Test Result for BERT and RoBERTa

Model	News Title	Result
BERT	Bill Clinton Hails Harris as 'President of Joy' but Warns Against Overconfidence	REAL 99.87% Confidence
	Bill Weld, Running as a Libertarian, Likens Donald Trump's Immigration Plan to Kristallnacht - The New York Times	FAKE 99.85% Confidence
RoBERTa	Bill Clinton Hails Harris as 'President of Joy' but Warns Against Overconfidence	REAL 95.35% Confidence
	Bill Weld, Running as a Libertarian, Likens Donald Trump's Immigration Plan to Kristallnacht - The New York Times	FAKE 99.9% Confidence

5. Results

As anticipated, it is observed that in all of the experiments conducted the RoBERTa model has outperformed the BERT model. The average accuracy received from all the experiments for BERT is about 95.95% and the average accuracy for RoBERTa stands at about 96.38%. Among all the experiments, it is noted that the training and validation losses are minimum along with highest the accuracy when the training batch size is 32 for both the models. However, when the weight decay values are altered, BERT demonstrates a better performance with a weight decay value of 0.1 as opposed to RoBERTa, which shows a greater performance with a weight decay value of 1.0 This shows that the training batch size of 32 is an optimum hyper parameter value along with a weight decay of 0.1 for BERT and 1.0 for RoBERTa.

The confusion matrices included for each of the experiments underscore the results observed in the above graphs that trace the training and validation loss for both the models. The graphs for all of the experiments show a decreasing trend in the training and validation loss over subsequent epochs. The inference time for the two models vary slightly, with RoBERTa taking more time to predict the label for a given news title. This time difference also indicates the complexity of the two models. The variation in inference time is mainly due to the additional

techniques made use by RoBERTa for classification, which also explains the higher accuracy rate. Both the models are able to accurately classify a news snippet that has not been included in the test data. The models' ability to accurately classify recent news as **FAKE** or **REAL**, implies a powerful solution to the widespread problem of dissemination of fake news.

6. Conclusion and Future Work

Therefore, it can be concluded that both RoBERTa and BERT models are fairly effective in tackling the problem of fake news detection. However, there is scope for future experiments and refinements. Currently, the models can only be used to classify news snippets in the English language. This has to be further expanded and researched upon to include various other languages for a global reach. Further experimentation has to be conducted as the same set of hyperparameters may not be applicable on a dataset containing news snippets in another language. Experimenting with newer and better models can also lead to a better performance and novel insights. Rather than training and testing on a static dataset, the model should be trained to work on live news feed. This will enable the model to base accurate classifications on dynamic real-world data.

Acknowledgments

The author wishes to express her gratitude to Professor Weiwei Sun, Associate Professor at the Department of Computer Science and Technology at Cambridge University, for her constant guidance and support. Additionally, this endeavour would not have been possible without the generous support of Xiaoxi Luo, a student at Peking University who served as the teaching assistant. This research has been conducted in collaboration with Cambridge Centre for International Research (CCIR). The author would like to thank the entire CCIR team for their continued support, editing help, and feedback.

References

- Ashish Vaswani, N. S., and Parmar, N. (2023). Attention is all you need. *arXiv – Computer Science*. <https://arxiv.org/pdf/1706.03762>.
- Baker, P. (2024). Bill Clinton hails Harris as ‘president of joy’ but warns against overconfidence. *The New York Times*. <https://www.nytimes.com/2024/08/21/us/politics/bill-clinton-dnc-speech.html>.
- Jacob Devlin, Ming-Wei Chang, K. L., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv – Computer Science*. <https://arxiv.org/pdf/1810.04805v2>.
- Kai-Chou Yang, T. N., and Kao, H.-Y. (2019). Fake news detection as natural language inference. *arXiv – Computer Science*. <https://arxiv.org/pdf/1907.07347v1>.

- Kai Shu, S. W., and Liu, H. (2017). Exploiting tri-relationship for fake news detection. *arXiv – Computer Science*. <https://arxiv.org/pdf/1712.07709v1>.
- Mudadla, S. (2023). Weight decay in deep learning. [Online]. <https://medium.com/@sujathamudadla1213/weight-decay-in-deep-learning-8fb8b5dd825c>.
- Rahali, A., and Akhlouf, M. A. (2021). Malbert: Using transformers for cybersecurity and malicious software detection. *arXiv – Computer Science*. <https://arxiv.org/pdf/2103.03806>.
- Samuel L. Smith, Pieter-Jan Kindermans, C. Y., and Le, Q. V. (2018). Don't decay the learning rate, increase the batch size. *International Conference on Learning Representations (ICLR)* - 2018. <https://arxiv.org/pdf/1711.00489>.
- Shrestha, A., and Spezzano, F. (2021). Textual characteristics of news title and body to detect fake news: A reproducibility study. *European Conference on Information Retrieval (ECIR)* 2021. <https://par.nsf.gov/servlets/purl/10230315>.
- Sinha, S. (2020). Fake news detection, data retrieved from Kaggle – *Fake News Detection*, <https://www.kaggle.com/datasets/samrat96/fake-news-detection/data>.

ARE GENERATIVE PRE-TRAINED TRANSFORMERS ALL WE CAN DO IF WE WANT TO HAVE A SMART MACHINE?

DAN CRISTEA^{1,2}, DANIELA GÎFU^{1,2},
ANDREI SCUTELNICU^{1,2}, CECILIA BOLEA¹

¹*Institute of Computer Science, Iași branch of the Romanian Academy*

²*Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași*

dan.cristea@acadiasi.ro,
{daniela.gifu,
cecilia.bolea}@iit.academiaromana-is.ro,
andrei.scutelnicu@info.uaic.ro

Abstract

Large Language Models (LLMs) have garnered widespread attention, as a leading-edge research domain in artificial intelligence (AI) and for their impressive real-world applications. The vectorised and neuronal basis of LLMs make them extremely skilful in mastering human language, excelling in language generation and contextual understanding, but less productive in tasks involving creative thinking, being criticised in the field of abstract reasoning, where they are said to produce logical inconsistencies or “hallucinations”. Some strengths and weaknesses of generative pre-trained transformer (GPT) models are analysed and discussed. The latest developments show enormous enhancements, which contradict many of the signalled weaknesses. We provide a bunch of examples showing both GPT’s good and poor performances on a number of tasks. A comprehensive search of the literature supports the major role of hybrid approaches in AI, integrating GPT solutions with classical methods based on structured knowledge, ontologies and symbolic reasoning – generally called symbolic AI.

Keywords: neuro-symbolic AI, generative pre-trained transformers, classical AI, ontology-based reasoning, mixed approach in AI, logical reasoning, deep learning, cognitive computing

1. Introduction

Large Language Models (LLM) enjoy today an extremely large audience, both as a very promising research field in AI and on the applications ground. The implementation of these models yielded extremely successful dialog systems, such as the range of ever better performing Open AI’s systems in the ChatGPT

family, Anthropic’s Claude, Meta’s LLaMa models and Eleuther AI’s GPT-J (Bommasani et al., 2023). The human-like dialogues they are able to engage in, more than producing a vivid enthusiasm in non-professionals, prove to be of real help in some domains, such as healthcare and medical diagnosis, content creation, education, legal advisory, where they are used to provide support (Lappin, 2024; Mohan et al., 2024). LLMs manifest amazing capabilities of language interpretation and generation, translation, or abstraction. They are trained to respond to prompts and are able to rephrase their produced output refining it repeatedly such that the user gets closer and closer to the kind of results they are expecting. These systems are able to memorise the dialogue and refer to entities occurring in the past prompted or generated text.

The literature is also prodigious in critics addressing these systems, but not all of them are well-documented. One of most pertinent discussions on the matter is due to Lappin (2024), who distinguishes between weak and strong arguments against LLMs. One of the critics Lappin dismantles says that LLMs do not add more to the basic data they are trained on combined with the user’s prompt (Bender et al., 2021; Chomsky et al., 2023a). His argument brings forward the subtle and sophisticated pattern identification and inference mechanisms of transformers, which permit them to excel in medical image analysis and diagnostics (Shamshad et al., 2023), same as revolutionising computational biology, by predicting properties of proteins and new molecular structures (Chandra et al., 2023), thus opening new ways for the development of new medications and clinical treatments.

Other voices remark the impossibility of LLMs to capture meaning, mainly because they are not grounded in real world experiences, as they use strictly **word** embeddings (Bender and Koller, 2020). The topology of meaning, however, can often lay on linguistic knowledge, or otherwise very few people would understand the idea of planets, viruses or underwater hunting. Moreover, the Multi-Model BERT (Lu et al., 2019) and GPT-4 (OpenAI, 2023), although being pre-trained on word embeddings, can identify or “understand” elements of a graphic image, respond to questions about them and even modify images.

A vivid debate regards to inner “mechanics” of language, by prolonging the classical views that language in humans is grounded on innate properties (Piatelli-Palmarini, 1980), which is not the case with GPT systems of today. Classical works in grammar formalisms (Chomsky et al., 2023b) regard grammars as the canonical form for expressing linguistic knowledge, positing the humans’ “language faculty” to encode a Universal Grammar, while the distributed representations of transformers do not possess such a symbolic level for representing grammar. Moreover, traditionalists also try to induce the discontent that Deep Neural Networks (DNNs) do not capture their language the way in

which humans achieve this knowledge. However, researchers like Smolensky (1987) and McClelland (2016), among others, consider that it is entirely possible that in humans language reside in non-symbolic, distributed representations, as it has been proved for some transformers capable to identify amounts of hierarchical syntactic structure and long-distance relations (Lappin, 2021; Hewitt and Manning, 2019), etc. There is also work that tries to incorporate syntactic tree structures into transformers like BERT, although the benefits do not yet seem to motivate the enriched model. Lappin argues that the adequateness of transforms to represent language is proved by their success in applications, rather than in the way in which they achieve these results. “From this perspective,” says Lappin, “criticising DNNs on the grounds they do not operate like humans is analogous to objecting to aircraft because they do not fly like birds.”

A strong argument against DNNs when facing them against humans is that the first requires much more data to learn than the second (Warstadt and Bowman, 2023). Also, recent results (Talman et al., 2021) prove that transformers like BERT do not learn inference through semantic relations between premises and conclusions. It is easily noticeable their incapacity to generalise inference rules through examples.

2. GPT Models

GPT models are Unsupervised Multitask Learners (Radford et al., 2019) known as transformer-based architectures, which handle sequential data types like plain language. They include in their architecture a stack of transformer-encoder layers and process text in a highly parallelised way. Pre-training is the fundamental idea underpinning GPT models, through which the model mainly acquires the statistical regularities of a language. The model gains an understanding of the underlying patterns, syntax, and contextual relationships in the text through this process. The model is refined on certain downstream tasks after pre-training, enabling it to perform well in a variety of NLP applications, including text production, translation, sentiment analysis, and more.

The OpenAI GPT family evolved from GPT-2 (released in 2021), including a transformer with 1.5 billion parameters, to GPT-3, with over 175 billion parameters, then to GPT-3.5, with almost 200 billion more parameters, and finally to GPT-4o (2022), with its astonishing 1.5 trillion parameters. This last version substantially improved GPT-3.5 and cemented its place as one of the most potent AI language models available at the time (Chude, 2023).

Their implemented paradigms of **zero-shot learning**, enabling the model to take on tasks for which it has not been explicitly taught, and **few-shot learning**, describing the model’s capacity to complete a task with only a few instances or

demonstrations, convincingly showed the model’s power to generalise and adapt to new tasks without requiring a lot of fine-tuning.

Among the distinguishing qualities of GPT-4o, can be remarked: a high reduction of training duration, a greater level of language comprehension, including an amelioration of reasoning skills, the capability to operate in multiple modalities, including text, graphics, and audio, and enhanced ethical and safety precautions for reducing biased behaviour, the spread of false information, and the creation of harmful content.

3. The Good News: LLMs Master Reliable Dialogues

In this section we will present proofs extracted from the latest variants of LLMs that are intended to clarify the level of interpretation and generation of natural language attained by these systems. We will restrict our study to only Open AI models, namely the ChatGPT models 3.5 and 4o.

The often-reported tendency of ChatGPT-like systems to generate biased or factually inaccurate content is usually referred to as “hallucinations”. The testers classify these failed language acts in mathematics (giving incorrect answers to prompts asking to perform computations), in requests for factual and common-sense knowledge (demonstrated weakness to possess an explicit knowledge about the world, to put in evidence or reason about the physics of objects and the usual behaviour of vivid entities), in situation tracking (the inability to maintain updated information about objects, agents, etc. in their dynamics), and in social reasoning (dysfunction to appreciate reactions of people in social interactions) (Mahowald et al., 2024).

However, our tests certify a contrary situation. In most of the cases the answers are coherent, proving a perfect “understanding” of the prompt, i.e. at a human-comparable level. We will give below a number of examples to prove this statement.

3.1. Mathematics

Me: *A have 11 cats and I bought two foxes. 3 cats ran away when they saw the foxes. This engaged a certain confusion on the foxes and one went out also. Finally, one cat returned. How many animals do I have now?*

GPT-4o: *Let's break it down step by step:*

- *You originally had 11 cats and 2 foxes, so that's a total of 13 animals.*
- *3 cats ran away, leaving you with 8 cats and 2 foxes (10 animals).*
- *1 fox also left, leaving you with 8 cats and 1 fox (9 animals).*
- *Then, 1 cat returned, so now you have 9 cats and 1 fox.*

In total, you now have 10 animals.

Our tests show that any arithmetics are now solved correctly, by accessing Python codes with adequate requests. The interesting solution here is that the model is capable of calling the right functions that solve the computations. This demonstrates that the model has overpassed the difficulty to apply stepwise problem-solving where precise logical reasoning is required, therefore when the answer isn't directly retrievable from pre-learned patterns. For instance, when prompting:

Me: *What is 1.5 to the power of 5?*

GPT-4o: *1.5 to the power of 5 is 7.59375.*

the Python program generated is:

```
# Let's calculate 1.5 to the power of 5
result = 1.5 ** 5
result
```

Which yields the result: 7.59375.

Another example:

Me: *Is 97,819 a prime number?*

GPT-4o: *No, 97,819 is not a prime number.*

And the inner call is:

```
# To check if 97,819 is a prime number, we can test
its divisibility by numbers up to its square root.
import sympy
# Check if 97819 is prime
is_prime = sympy.isprime(97819)
is_prime
```

More elaborate inner computations are hidden to the user.

3.2. Factual Information

Me: *What was the famous music festival that went on in Bucharest in September this year?*

GPT-4o: *The famous music festival that took place in Bucharest in September 2024 was the **George Enescu Festival**. This prestigious international classical music event is one of the largest in Romania and attracts orchestras and performers from all over the world. The 2024 edition ran from **August 31st to***

September 27th, and it featured numerous performances by renowned artists and orchestras in venues like the Romanian Athenaeum.

The festival is dedicated to the great Romanian composer George Enescu and is a highlight in the classical music world.

It also reports having searched 4 sites for information.

Another example:

Me: *Who was the U.S. President during World War I?*

GPT-4o: *The U.S. President during World War I was Woodrow Wilson. He served as the 28th president of the United States from 1913 to 1921. Wilson led the country through the war, which lasted from 1914 to 1918, with the U.S. entering the conflict in 1917.*

The straightforward solution was to make use of existing knowledge on the web, by first finding the right pages which include the needed information and then abstracting the information found there, very much on the line with the LLMs' inner paradigm.

3.3. Conceptualised Knowledge of Syntax

Me: *Is it an error in the sentence: "The key to the cabinets are on the table."?*

GPT-4o: *Yes, there is an error in the sentence "**The key to the cabinets are on the table.**" The subject of the sentence, "the key," is singular, but the verb "are" is plural. To correct the sentence, the verb should agree with the singular subject. The corrected version is:*

"The key to the cabinets is on the table."

Here, "is" correctly agrees with the singular subject "key." The phrase "to the cabinets" is a prepositional phrase, which doesn't affect the number of the subject.

The dialogue proves not only a correct mastering of the syntax of language, which is an acquired benefit of training on a large set of texts, but also the conceptualisation of this knowledge. Within the LLM, this level of abstraction could be acquired only by an attentive labelling of the syntactic roles in sentences, a possibility which has been demonstrated in some previous work (Hewitt and Manning, 2019). One possibility to do this is through Recursive Neural Networks (RNNs), particularly the Tree-Structured RNNs (Socher et al., 2011), which explicitly model tree-like syntactic structures. These networks parse sentences into constituent parts and recursively apply neural operations at each node, passing up the updated tree and maintaining hierarchical relationships between words. This type of processing can equally apply to semantic relationships,

making it particularly well-suited for tasks like semantic parsing or syntax-aware sentiment analysis.

3.4. Object Permanence and Physical World Consistency

Me: *If I put a book in a box and then close the box, and after that I open the box, where is the book?*

GPT-4o: *If you put a book in a box, close the box, and then open it again, the book will still be inside the box, right where you placed it—unless someone or something has moved it in the meantime. The location of the book is unchanged by simply closing and reopening the box.*

Me: *Maria is cooking in the kitchen. After dropping an egg from her hand, what do you think Maria will do?*

GPT-4o: *If Maria drops an egg while cooking, it's likely that she will clean up the mess first—wiping up the broken egg and possibly cleaning the floor or counter. After that, she may get another egg and continue cooking, depending on what she's preparing.*

Me: *If you drop a glass on the floor, what happens?*

GPT-4o: *If you drop a glass on the floor, it will most likely shatter or break into pieces, depending on the height it was dropped from and the type of floor. A hard surface like tile or concrete is more likely to cause the glass to break, while a softer surface like carpet might cushion the fall and prevent it from shattering.*

Issues of this kind are generally grouped under the general term of common-sense. The acquisition of common-sense in machines has fired the minds of AI researchers for a long time. Usually, humans do that by their day-by-day interaction with the exterior world. From the child's age till reaching maturity, humans learn how to interact with the environment, thus experiencing its laws, mainly mastered by physics, helping them to make correct deductions, even in situations which hide details. Common-sense reasoning has extremely many facets and one or two examples cannot certify that LLMs really master the diversity of the world. However, it seems that linguistic knowledge, abundantly fed into these systems during training, also contributes with convincing common-sense knowledge to allow coherent dialogues on the matter. If this has solved the much debated “AI bottleneck” problem is still an open question.

3.5. Spatial Awareness and Positional Reasoning

Me: *If I have a table with a chair on the right and an armchair on the left and I sit on the chair is the table to my right or left side.*

GPT-4o: *If you sit on the chair, which is on the **right** side of the table, then the table will be to your **left** side.*

Me: *"If you place a cup on a table and then sit on the floor, is the cup above or below you?"*

GPT-4o: *If you place a cup on a table and then sit on the floor, the cup is above you. Since the table is at a higher level than the floor, and you're sitting lower than the table, the cup is positioned above you.*

Not long-ago, questions involving spatial relations were an area where GPT models faltered. The correct interpretation, as proved in the chats above, stems from the model's training on real-world awareness and its new acquired ability to reason about object positions logically. However, it is questionable if this awareness can be obtained solely by linguistic training without the incorporation of a sophisticated spatial model.

3.6. Situations Tracking

Me: *A cat is on a horse. They are riding. The horse falls down. Where is the cat?*

GPT-4o: *If the horse falls down while the cat is on it, the cat could be in several possible places depending on how it reacts. Cats are typically agile and quick to react, so most likely, the cat would jump off before or as the horse falls. The cat could end up standing on the ground nearby, having leaped off to safety. If the cat didn't have time to react, it might end up on the horse or near the horse on the ground. Cats tend to land on their feet, so it would probably be unharmed.*

This issue refers to the ability of humans and LLMs to foresee the development of a story based on general considerations about the usual behaviour of living beings. The difficulty comes from at least two directions: extracting information from many sentences in a row, which obliges for a continuous increase of the models' context window, i.e., the number of words they can process in one go (a model that simultaneously attends to each word in a book is vastly inefficient) and integrating incoming inputs to dynamically update information about entities and their states, which makes necessary to maintain a database of positions and moments. However, generalised multi-modal DNNs achieve this competency by pairing a suitable graphic (and other modes of) representation with a sentence that describes a situation (Kim and Schuster, 2023).

3.7. Theory of Mind

Me: *John put the toy in the box and left. Mary secretly moved it in the wardrobe. Where does John think his toy is?*

GPT-4o: *This scenario is a classic example used to illustrate the concept of **false belief** in developmental psychology, often called a “Theory of Mind” experiment. Here’s how it works:*

*Since **John put the toy in the box and then left**, he believes the toy is still in the box because he hasn’t witnessed **Mary secretly moving it to the wardrobe**.*

*In this case, John’s belief is based on the information he had before leaving, so **John thinks his toy is still in the box**.*

This test demonstrates the ability to understand that other people can hold beliefs different from reality or from your own, which is a key aspect of developing Theory of Mind (ToM), typically around the age of four-five years in children.

Theory of Mind is defined as the cognitive ability to predict and plan people’s actions by understanding their perceptions, beliefs, and desires. Mahowald et al. (2024) comment that LLMs’ ability to solve ToM tasks, which has been subject to particular controversy, require both social knowledge and the ability to maintain a situation model. Contrary to many pessimistic opinions (Gandhi et al., 2023), ChatGPT-4o seems to master at least some of the necessary power to handle ToM. The supposition is that LLM is augmented with a symbolic tracker of entity states and character beliefs (Sclar et al., 2023). People also speculate that this ability may occur naturally and independently as a consequence of the existence of undisclosed regularities in language that enable the resolution of ToM-like tasks without the direct engagement of ToM (Kosinski, 2023).

3.8. Human-Like Social Interaction

Me: *If I smile at a stranger, what might they do?*

GPT-4o: *If you smile at a stranger, they might:*

1. **Smile back** as a friendly or polite gesture.
2. **Nod or acknowledge** you in some way, like with a brief greeting.
3. **Ignore the smile** or not notice it, depending on their focus or mood.
4. **Feel confused or surprised**, especially if they aren’t used to random friendly gestures.

In many cases, a smile is seen as a positive, welcoming action, so the stranger is likely to respond warmly.

GPT models were accused of displaying illogical responses in social reasoning scenarios. However, this test reflects how the model perfectly understands common human behaviour, providing human-like social reactions. This is a proof that LLMs can acquire this type of knowledge with features that mimic human performance in social interactions. It is still questionable if this can be obtained purely through linguistic training. The models’ ability to engage in truly deep

emotional intelligence is still quite limited compared to human interactions. Intense studies go in the direction of gaze identification and recognition of emotions in videos in connection with language (Hu et al., 2022). Also, there are new studies in densely aligned language-visual representations and acquisition of patterns from these resources (Sangmin et al., 2024).

3.9. Summarising Long Texts

Me: *If you read Tash Aw's book "The map of the invisible world", can you tell me: at the moment when Adam meets Margaret was Karl around or not?*

GPT-4o: *In Tash Aw's "The Map of the Invisible World", when Adam meets Margaret, **Karl** is not around. Karl, who is Adam's adoptive brother, has disappeared, which is a central element of the story. Adam is searching for Karl, and Margaret, an older woman who knew Karl in the past, becomes someone Adam seeks out for help. At the moment of their meeting, Karl remains missing, which adds to the tension and mystery surrounding his whereabouts.*

Me: *But are you sure that Karl is Adam's adoptive brother?*

GPT-4o: *You're correct to question that—**Karl is not Adam's adoptive brother**, but rather his **adoptive father** in "The Map of the Invisible World" by Tash Aw. Karl, a European expatriate living in Indonesia, adopts Adam when he is a child. The novel revolves around Adam's search for Karl, who is arrested by the authorities, leading Adam to meet Margaret, who had a past connection with Karl.*

So, to clarify, when Adam meets Margaret, Karl (his adoptive father) is not present, as he has already been taken away.

It has long been argued that LLMs are not able to keep track of facts related in a text at distance. In this example we bring in discussion a novel which is longer than 400 pages. GPT-4o is able to find relations about characters even if the events they are involved in are extremely distant in the text. A minor mistake is immediately corrected in this example. Time processing in texts is an issue of intense study in classical AI and NLP. Time theories (Allen, 1983; Pustejovsky et al., 2003) formalise the way humans understand assertions involving time and how can time be annotated and processed. A solid solution to this issue is to integrate time reasoning models into LLM (Cristea, 2017).

3.10. Anaphora Resolution

Me: *If the baby does not thrive on raw milk, boil it. Whom "it" refers to?*

GPT-4o: *In the sentence "If the baby does not thrive on milk, boil it," the word "it" technically refers to **the milk**. However, the sentence is confusing because "it" could easily be misinterpreted as referring to the baby. The intended meaning*

*is that the **milk** should be boiled, but the wording should be clarified for better understanding, like:*

- *"If the baby does not thrive on milk, boil the milk."*

This removes any ambiguity.

Anaphora resolution (AR) is known as a fertile field of study in natural language processing. Among researchers working in this field, circulates the truism that anaphora resolution is offered to you for free the very moment you solve everything else. Many examples have been invented to show the difficulty of automatic linking references to antecedents mentioned in the text. In many cases these are long distance links and involve knowledge of syntax, semantics and pragmatics to disambiguate. The reproduced joke is well known in the AR community. So, it seems that LLMs are able to also latch on pragmatic knowledge, of the kind to make a clear distinction between what can be boiled and what not between a baby and the milk. Frequent linguistic associations can incorporate this type of experience, which humans acquire mainly by a mere contact with reality.

3.11. Responding to Incoherent Prompts

Me: *Sally doesn't own a dog. The dog is black.*

GPT-4o: *It seems there may be a small contradiction in your statement: if Sally doesn't own a dog, how can the dog be black? Could you clarify what you meant, or if you're referring to some other situation or example? I'd be happy to help clarify further!*

Me: *Am I wrong or you really trill?*

GPT-3: *Since I don't actually speak out loud, I don't produce any vocal sounds such as trilling or burring. My responses are generated as text. However, if I were capable of speech, I would ensure proper pronunciation according to the language I'm communicating in. 😊*

Me: *Cramming the hinges over the bald princess's blonde hair!*

GPT-4o: *That sentence sounds quite surreal and abstract! It mixes different, unrelated elements – "cramming hinges," "bald princess," and "blonde hair" – creating an image that's hard to visualize logically. It seems like a playful or poetic use of language, perhaps meant to evoke confusion or provoke thought.*

If you're aiming for a surreal effect, it certainly achieves that! Let me know if there's a particular context or meaning you're working towards with this sentence.

Above, there are three attempts to confuse the system with prompts which communicate nonsense assertions. However, the model correctly handles them. The oddness of situations communicated in language is mastered well and a

nonsensical sentence is brought to light and signalled. There is no trace of hallucination in its responses. We are, again, on the realm of acquisition contributed as a side effect of feeding the system, during training, with a huge amount of textual data.

4. But There Are Also Bad News: LLMs' Weaknesses

4.1. Dealing with Inferences

The inability of LLMs to make deductions while they struggle with complex reasoning tasks is the most frequent discontent of LLMs critics (Lin et al., 2024; Lappin, 2024). We fabricated the small problem below to test this feature:

Me: *John and Mary have 4 children: Luisa, Bazil, Sorin and Mircea. Only one of their boys has a sister who is older than himself and only Mircea has a younger brother. Sorin is younger than Bazil. Who is the firstborn?*

Our reasoning:

IF *only one of John and Mary's boys has a sister who is older than himself*, THEN Luisa, being the only girl, should be in the 3rd position. Therefore, the positioning from the first born (the older) to the last one (the younger) should be *boy boy Luisa boy*, because this is the only case in which Luisa could have one and only one younger brother.

IF *only Mircea has a smaller brother*, THEN Mircea should be in the second position, like this: *boy Mircea Luisa boy*.

IF *Sorin is younger than Bazil*, THEN it is clear the order of the last two: *Bazil Mircea Luisa Sorin*, therefore Bazil is the first born.

Let's see now for ChatGPT-4o's way?

GPT-4o: *Let's break this down step by step with the clues provided:*

1. **Only one of the boys has a sister who is older than himself.**
 - *This means one of the boys has **Luisa** (the only sister) older than him, but not all of them.*
2. **Only Mircea has a younger brother.**
 - *This means Mircea cannot be the oldest or the second oldest; he must have at least one younger brother. Therefore, Mircea must be **second to last** (third child), and the fourth child is his younger brother.*

3. *Sorin is younger than Basil.*

- *This tells us the relative ages of Sorin and Basil. Since Sorin is younger than Basil and Mircea is third, Sorin must be the fourth child, and Basil must be second.*

Let's summarise the order:

- *Luisa is the only sister, and based on the first clue, she must be the eldest.*
- *Basil is the second child.*
- *Mircea is the third child.*
- *Sorin is the fourth child (the youngest).*

Conclusion: *Luisa is the first born.*

When we signalled the error, the dialogue went on, every time with a wrong answer. Then, we prompted GPT why it is wrong:

Me: *Let me explain to you. If only one of their boys has a sister who is older than himself then it means that the sister, i.e. Luisa, should be in the third position, because only this way she has only one brother younger than herself. Now, can you go on?...*

With this clarification it finally got the correct order and the system indicated the firstborn.

This problem is clearly one of constraints satisfaction. The domain of AI has dealt with this type of problems for a long time and solutions have been elaborated. Other well-known problems of this kind are: the *8 queens problem* (how to place 8 queens on a chessboard such that no queen should be attacked by no other queen), the *map-colouring problem* (having a map of n countries, assign a colour to each country such that any two neighbour countries to have different colours), the *Sudoku game* (starting from a partially filled 9x9 board, place a figure from 1 to 9 on each position so that each one appears only once on a line, on a column and in a 3x3 region). In general, constraint satisfaction problems are intractable (NP-hard). The techniques used in solving some of such problems are divided into two large categories: search and inference. The search algorithms traverse the space of partial solutions building a complete instantiation that satisfies all constraints, or reports the inconsistency of the problem. Backtracking schemes are part of these techniques. Consistency inference algorithms modify the problem at each step, in order to make it more explicit, therefore easier to solve. There are also algorithms that combine the two methods, prone to more accurate results (Cristea et al., 2007).

It has not been tested enough the ability of transformers to embrace inference abilities at a level which could be considered more than superficial. To accommodate this type of reasoning, the training corpora would have to include labelled patterns and in a large quantity. In their very recent study, Mahowald et al. (2024) review substantial amounts of evidence for the claim that LLMs do not perform well on extra-linguistic reasoning and real-world knowledge tasks. They make the distinction between formal linguistic competence, as “the knowledge of rules and statistical regularities of language, and functional linguistic competence - the ability to use language in real-world situations”.

Mahowald et al. (2024) dismantle the false supposition that being good at language should necessarily imply being good at thought, which is clearly not the case neither for humans nor, especially, for LLM-based models of today. There is no surprise in this discrepancy if we notice that models based on the “guess next word” objective can easily acquire formal linguistic competence but not necessarily functional linguistic competence. The inherent suction of a part of the world's knowledge that comes together with the huge amount of language served to these models during training could also stabilise the (false) impression that these systems are good at making reasoning and at achieving context-dependent goals.

A comprehensive review of the recent literature about reasoning with foundation models (LLMs) can be found in (Sun et al., 2024).

4.2. Counting Similar Items

Me: *How many “r” are in Raspberry?*

GPT-3: *The word “Raspberry” contains **three “r”s**:*

- *One at the beginning: **R**aspberry*
- *One in the middle: Rasp**er**ry*
- *One at the end: Raspber**rr***

So, the total count is 3.

Me: *How many “r” are in Raspberry?*

GPT-4o: *The word “Raspberry” contains **two “r”s**. One “r” appears at the beginning, and the other is located in the middle of the word.*

Me: *How many stars are on this line: * * * * * ?*

GPT-4o: *The line “* * * * * ” contains **9 stars**.*

Me: *How about now: *gx*t*(*rrR*)?*

GPT-4o: *The sequence “gxt*(rrR)” contains **4 stars**.*

It is rather curious this inability of GPT to properly count and to correctly reproduce the inputted strings. We have no plausible explanation other than no special training of these types of tasks triggers bad functioning.

5. Conclusions

In this paper we proposed an analysis of how the current GPT models in the OpenAI family engage in tasks which, till recently, were considered not properly covered by their competences. These tasks involve processing which is not comfortably handled by LLMs. Many of these tasks however have solutions developed with traditional, symbolic AI methods, like rule-based inference, search in the state-space, planning, conceptual representations and ontology-based reasoning, or NLP methods like anaphora resolution, syntactic and semantic parsing, discourse understanding and generation. These symbolic methods have been, time ago, successful in domains like theorem proving, robotics, expert systems, question-answering, summarisation, sentiment analysis, etc., but their ability to generalise across different tasks proved to be limited compared to the probabilistic and neural network-based architectures used in LLMs. Despite this, LLMs' flexibility introduces issues such as dependency on large, static datasets and difficulties in handling real-time updates, which are critical in domains that rely on the latest information¹, such as medical diagnosis and legal services (Bharathi et al., 2024).

Only a rough evaluation of the performances of LLM systems is dared in this paper. Methodologies for evaluating LLMs, pinpointing specific areas and aspects for assessment, exploring their contextual appropriateness, are vital for gauging LLM efficacy. Their stability and reliability when encountering unexpected or challenging inputs and their effective adaptation in specialized sectors, including healthcare, finance, and law, makes necessary a firm estimation of the performances and stability of these systems. LLM evaluation has overpassed already its infancy, since a whole range of benchmarks have already appeared. In a recent post, Abose (2024) inventories 27 LLM evaluation benchmarks, and the area is estimated to develop intensively.

The often-reported weaknesses of LLMs sparked new directions of investigation, making developers shift away from scaling-up the language-prediction ability to more specialised methods targeting reasoning skills. Among these are *Reinforcement Learning from Human Feedback* (RLHF) (Ouyang et al., 2022) or the coupling of LLMs with external specialised modules, leading to so-called *augmented language models* (Mialon et al., 2023) or, very

¹ <https://www.projectpro.io/article/llm-limitations/1045>

recently, *agentic AI*². Also, methods such as scheduled retraining and incremental learning are being developed to mitigate LLMs' static nature, enabling them to update their knowledge without requiring full retraining. These developments configure the new range of neuro-symbolic AI, which seems to gain confidence in present days.

References

- Brain John Aboze. 2024. How to Measure LLM Performance, <https://www.deepchecks.com/how-to-measure-llm-performance/>.
- Allen, James F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11).
- Emily M. Bender and Alexander Koller. 2020. Climbing Towards NLU: On Meaning, Form, and Understanding in the Age of Data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5185–5198, DOI: 10.18653/v1/2020.acl-main.463.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? *FACCT '21, Association for Computing Machinery*, pages 610–623.
- Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. 2023. The foundation model transparency index, arXiv preprint arXiv:2310.12941.
- Abel Chandra, Laura Tünnemann, Tommy Löfstedt, and Regina Gratz. 2023. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12.
- Naom Chomsky, Ian Roberts, and Jeffrey Watumull. 2023a. The false promise of chatgpt. *The New York Times*, March, 8.
- Naom Chomsky, R. C. Berwick, and G. Marcus. 2023b. Linguistic Fundamentals in AI: Can Machines Truly Understand Human Language? *Journal of Cognitive Science*, 47(1):7-25.
- Emmanuel Chude. 2023. GPT-3.5 and GPT-4 Comparison: Exploring the Developments in AI-Language Models, August 4, online at <https://medium.com/@chudeemmanuel3/gpt-3-5-and-gpt-4-comparison-47d837de2226>.
- Dan Cristea, Mădălina Ioniță, and Ionuț Cristian Pistol. 2007. *Inteligență Artificială (EN: Artificial Intelligence)*, Editura Universității “Alexandru Ioan Cuza”, Iași.
- Dan Cristea. 2017. The Time Yards Model: Rethinking the way to look at texts. *Proceedings of SYNASC-2016*.
- Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding social reasoning in language models with language models. *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, <https://arxiv.org/abs/2306.15448>.
- John Hewitt and Christopher Manning. 2019. A structural probe for finding syntax in word representations. *Proceedings of the 2019 Conference of the North American Chapter of*

² See a survey at <https://www.analyticsvidhya.com/blog/2024/12/ai-agents-research-papers/>

- the Association for Computational Linguistics: Human Language Technologies*, 1: 4129–4138. Association for Computational Linguistics, Minneapolis, Minnesota.
- Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. 2022. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Kim, N., Schuster, S. 2023. Entity Tracking in Language Models, <https://arxiv.org/pdf/2305.02363>.
- Kosinski M. 2023. Theory of mind may have spontaneously emerged in large language models. <https://arxiv.org/vc/arxiv/papers/2302/2302.02083v1.pdf>
- Shalom Lappin. 2024. Assessing the Strengths and Weaknesses of Large Language Models. *Journal of Logic, Language and Information*, 33, 9–20, <https://doi.org/10.1007/s10849-023-09409-x>.
- Shalom Lappin. 2021. *Deep Learning and Linguistic Representation*. CRC Press.
- Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, James M. Rehg. 2024. Modeling Multimodal Social Interactions: New Challenges and Baselines with Densely Aligned Representations, <https://arxiv.org/html/2403.02090v1>.
- Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. 2024. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57:243.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13–23.
- Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models: a cognitive perspective. [arXiv:2301.06627 \[cs.CL\]](https://arxiv.org/abs/2301.06627).
- James L. McClelland. 2016. Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. In B. MacWhinney and W. O’Grady (Eds.), *The Handbook of Language Emergence*, pages 53–80, John Wiley and Sons.
- Grégoire Mialon, Roberto Dessì, Maria Lomeli, et al. 2023. *Augmented language models: a survey*, [arXiv preprint arXiv:2302.07842](https://arxiv.org/abs/2302.07842).
- G. Bharathi Mohan, R. Prasanna Kumar, P. Vishal Krishh, et al. 2024. An analysis of large language models: their impact and potential applications. *Knowledge and Information Systems*, 66:5047–5070.
- OpenAI. 2023. GPT-4 technical report, [arXiv:2303.08774](https://arxiv.org/abs/2303.08774).
- Long Ouyang, Jeff Wu, Xu Jiang, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- M. Piatelli-Palmarini (Ed.). 1980. *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*. Routledge and Kegan Paul, London, pp. 409.
- Pustejovsky James, Castano José M., Ingria Robert, Sauri Roser, Gaizauskas Robert, Setzer Andrea, Katz Graham, Radev Dragomir R. 2003. TimeML: Robust specification of event and temporal expressions in text. AAI Technical Report SS-03.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners, Open AI blog,

- https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
- Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding language models’ (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 1: 13960–13980, Toronto, Canada, July.
- Fahad Shamshad, Salman Khan, Syed Waqas Zamir, et al. 2023. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88.
- Richard Socher, Cliff Chiung-Yu Lin, Andrew Y. Ng, and Christopher D. Manning. 2011. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 129-136.
- P. Smolensky, 1987. Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review*, 1(2): 95–109.
- Sun et al. 2024. A Survey of Reasoning with Foundation Models, <https://arxiv.org/abs/2312.11562>
- Aarne Talman, Marianna Apidianaki, Stergios Chatzikyriakidis, and Jörg Tiedemann. 2021. NLI data sanity check: Assessing the effect of data corruption on model performance. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 276–287, Linköping University Electronic Press, Sweden, Reykjavik, Iceland (Online).
- Alex Warstadt and Samuel R. Bowman, 2023. What artificial neural networks can tell us about human language acquisition, *Algebraic Structures in Natural Language*, pages 17–59, CRC Press.

THE DISCOURSE OF PROFESSIONAL COMMUNICATION WITH YOUNG LEARNERS: CAREER ORIENTATION PRACTICES IN THE FIELD OF COMPUTATIONAL LINGUISTICS

ANDREEA NECHIFOR¹, MARIA UNGUREANU^{2,3}

¹ *'Transilvania' University, Braşov*
andreeabratan@unitbv.ro

² *Doctoral School, State University of Moldova*

³ *Theoretical College 'Miron Costin' Paşcani*
mariamarrynna@gmail.com

Abstract

This study investigates the significance of work in shaping individual identity and the role of effective career guidance in preparing young people for successful transitions into the workforce. It explores the evolving field of Computational Linguistics, highlighting the interdisciplinary nature of this domain and its growing importance in the digital age. The study examines a successful collaboration between ConsILR and high schools in the Today's Student, Tomorrow's Researcher initiative, demonstrating the potential of such partnerships to inspire young people to pursue careers in research and contribute to advancements in this field. The study concludes by acknowledging the indirect but significant contributions of young people to the field of Computational Linguistics through not only the provision of valuable linguistic data, but also direct research input. Additionally, the study highlights the importance of early exposure to diverse career paths, such as through school and vocational orientation programmes, summer schools, and interactions with professionals in various fields. The 19th edition of the ConsILR conference, which included a special section for high school students, serves as an example of how such initiatives can introduce young people to the exciting world of Computational Linguistics and inspire them to pursue research in this field.

Keywords: career guidance, Computational Linguistics, work identity, youth engagement, labour market, professional communication

1. Introduction

Since ancient times, hardworking individuals have been valued and respected in any human community. Statements such as 'work ennobles' or 'work is second only to prayer' express the importance humanity has attributed to labour,

recognizing its potential to shape individuals both morally and socially. In its broadest sense, defined on philosophical grounds, work can be understood as the activity through which humans produce both material and spiritual goods. In a narrower sense, according to modern perspectives, work refers to the activity of producing socially recognized goods and services, serving also as a source of income for those who perform it (Zamfir and Vlăsceanu, 1998: 372-373). From a sociological perspective, work is defined as the fulfilment of tasks necessary to ensure existence within a specific social environment (Watson, 2003: 113). However, the concept of ‘earning a living’ transcends the mere production of material goods intended to ensure physical survival. Thus, people do not simply extract their means of subsistence from the environment in which they live; through their activities, they transform the surrounding environment, creating living conditions that go beyond the minimum necessary for survival. In this sense, work becomes a determining factor not only for meeting material needs but also for shaping and developing an individual’s personal identity.

Initially cultivated and subsequently invested in as a defining factor in determining future careers, professional profiling has lately been developed as a ground-setting factor by educational professionals in the academic world. Specialized vice-rectorates have been appointed and designated persons have been involved in creating the most appropriate and attractive events and projects, as well as in implementing institutional policies to make the undergraduate world become prepared for the economic market (Munteanu and Nechifor, 2019, 2023). But in a world dominated by capitalism and mercantilism set against the background of very fast-paced technological advancements, as it is the case of reality nowadays, shaping identities and future paths should start at much younger age, i.e., high-school, if not even before. Selecting a specific profile at the end of eighth grade to follow for the next four years has increasingly proven to be insufficient. Greater emphasis should be placed on the particular subjects to be studied, allowing for a more personalized approach that aligns with students’ evolving interests. In this context, enabling individual choices becomes essential (Purcaru and Nechifor, 2013). Events such as academic conferences, which serve as a bridge between education and future career pathways, provide an ideal platform for high school students to explore their interests and better prepare for the labour market.

2. School and Vocational Guidance

The perception of work varies from one community to another, and from one individual to another, as it results from the interplay of several factors that contribute to shaping and consolidating it: political climate, the moral and

pragmatic values of the family environment, the characteristics of the social context in which the individual is integrated, education, innate abilities, role models, mentors, and more. A recent study provided by Nechifor and Dimulescu (2024) proves that even the choice of using or not different digital tools can be influenced with students by a wide range of factors like these, but the major advantage is that as a result of this, they are if not equipped with the necessary digital skills so necessary in the modern world of today, at least exposed to them in order to be prepared to the requirements of the future employers.

The shaping of a young person's perception of work begins in the early years of life and continues, intensifying, during the school years, through methods specific to each educational level. Thus, in primary school, pupils are guided to explore the vast range of professions, primarily from the perspective of their importance in society. In middle school, the educational process is materialized in activities that facilitate self-awareness. Pupils are encouraged to recognize their skills and to understand their interests concerning a future career. At the high school level, during both classes and career orientation sessions, pupils are provided with the necessary information in order to be able to choose a field of activity compatible with their personality traits and personal aspirations.

This stage of their academic journey is particularly decisive, as it influences their ability to find fulfilment in the profession they will ultimately choose. The success of a future specialist in a given field depends substantially on how they were guided and the level of development of their perception of work.

According to Gati and Asher (2001:491), school and career orientation is “a psychological and educational process aimed at supporting individuals in clarifying educational and occupational choices by integrating self-knowledge with environmental information”. In other words, school and career orientation are two essential components of the educational and vocational counselling process, each with a specific focus and distinct contributions to personal development and the socio-professional integration of the individual. Although closely interconnected, they differ in their objectives, activities, and goals.

School orientation primarily aims to guide students in choosing an educational path that aligns with their interests, aptitudes, and personal values, which are essential for making informed educational and career decisions (Savickas, 2013). This process focuses on cultivating academic success and ensuring “smooth transitions between various educational stages” such as either from middle school to high school or, more responsibly, from high school to university. Specific activities include administering aptitude and vocational orientation tests, providing counselling to select an appropriate educational track (humanities, sciences, technology, etc.), and offering information about the specific requirements of educational institutions. Students are informed about the available

educational pathways, professional requirements, and opportunities in various fields (Watts and Sultana, 2004).

In contrast, career orientation focuses on preparing individuals for integration into the labour market and the development of a career. This type of orientation aims to identify vocational interests and transferable skills, supporting young people and adults in making decisions regarding their professional path. Activities include counselling on career selection or professional training programmes, analysing labour market requirements and trends, and preparing for employment processes, such as writing a CV or preparing for interviews. According to Super (1980), career orientation helps individuals identify and develop appropriate professional roles, taking into account their personal values and the existing economic opportunities. Career orientation programmes develop skills such as career planning, time management, and option evaluation to enable well-informed decision-making (Brown and Lent, 2013).

In summary, school orientation is directed toward educational success and effective transitions between educational cycles, while career orientation focuses on labour market integration and career planning. The two processes are complementary, supporting individuals in making well-grounded decisions in both educational and professional domains. Together, they contribute to shaping autonomous individuals capable of adapting to the dynamic demands of modern society and the economy.

The decision to choose a profession is influenced by a multitude of factors from both school and extracurricular environments. In this context, close collaboration between an educational establishment and all actors involved in the career orientation process becomes essential. The school body plays a significant role, utilizing numerous tools to support pupils and students in making informed decisions about their professional future. By its institutional nature, the school can centralize and provide students with access to a wide range of documentary resources useful for exploring various professions. These resources include professional monographs and micro-monographs, informational boards, thematic folders about occupational fields, as well as materials dedicated to the educational network, such as guides, brochures, or information on admission competitions. These tools contribute to effective and well-founded student orientation.

However, consulting such resources is not enough, as the information they provide is general. When choosing a profession, young people need to go beyond the surface image of a given field and access an insider's perspective. This can be offered by experts in the fields that interest the student. Favourable contexts for this type of interaction include short-term internships or student visits to companies and specialized institutions, activities which are usually conducted under the guidance of experts, but also by conferences, and more. Through direct

dialogue with experts, students can obtain concrete and reliable information aligned with their interests, which will serve as the foundation for their decision about the career they wish to pursue.

At a pre-university level, a nationally devised plan such as ‘*Școala Altfel*’ (EN: *School Otherwise*) can provide pupils with an alternative non-formal exposure to such informed and practical decisions through a programme that contributes to diversifying students’ educational experiences, offering them the opportunity to learn in a way different from traditional classroom lessons. ‘*Școala Altfel*’ has as its main objectives, according to OM 29.08.2016, the promotion of learning through alternative and interactive methods, the development of creativity and team spirit among students, and the active involvement of students in activities designed to stimulate their interest in knowledge and personal development. This programme encourages active participation, creativity, and the development of essential practical skills for their personal and professional growth. As part of ‘*Școala Altfel*’, educational institutions organize various extracurricular and non-formal activities tailored to the age and interests of students, such as: visits to museums, theatres, cultural or scientific institutions, workshops on artistic, scientific, or technological creation, community and volunteer projects, competitions and contests on various topics, and sports and recreational activities.

At a university level, internship represents the pragmatic means via which students can make informed decisions about their future career, as a result of having a first-hand experience regarding the potential nature of the jobs they envision for themselves after graduation. Combined with dedicated subject-matters, such as Professional Communication Strategies, which analyse, depending on the study programme, the characteristics of the affiliated jobs students may choose as a result of their profiles, internship stands as the first step towards correlating didactic and research activity with the economic and business environment, in other words, between the offer and the needs. And one particular way in which, according to Munteanu and Nechifor (2024), as a next step after internship, universities can develop this into the student population is through involving them into intrapreneurial activities within the educational establishment they belong to, with mutual benefits: assisting students’ ability to develop their entrepreneurial skills while developing the academic institution through the original ideas and projects put forward by the same students. An example in point is “Transilvania” University of Brasov, where projects such as: *10 Ideas from 20.000 Students*, *My Faculty*, *Together for University*, *We Develop the University*, and, more recently, *Learn to Be Central!* stand as proofs of students’ intrapreneurial potential. Student Entrepreneurial Society (SAS-UNITBV) in collaboration with The Entrepreneurs’ Club (ENC) created a for 40-hour UNITBV Green – Life HACKATHON, event which enabled young participants to present

their ideas, strengthen their teams and develop a prototype for the supported concepts, getting advice from mentors (entrepreneurs – members of ENC). Last, but not least, a direct meeting and interaction between companies and undergraduates takes place every year, on the occasion of the event called *Absolvenți în Fața Companiilor – AFCO* (EN: *Students Meet Companies*), which provides senior students of the bachelor and master programmes with the opportunity to present their graduation projects to potential employers and discuss the real needs of the labour market.

Professional communication strategies have emerged as a cornerstone of undergraduate education across disciplines, with studies from major universities highlighting its critical role in workforce preparation. Research published in the *Journal of Business and Technical Communication* (Rahman et al., 2023) suggests that students who complete dedicated professional communication courses demonstrate a 37% higher rate of successful job placement within six months of graduation compared to those without such training. The field encompasses both theoretical frameworks and practical applications, ranging from interpersonal communication theories to digital media competencies, with particular emphasis on cross-cultural communication in an increasingly globalised workplace. A comprehensive study conducted by the *National Association of Colleges and Employers* (NACE, 2022) identified effective communication as the most sought-after skill by employers, with 91% of hiring managers rating it as ‘very important’ or ‘essential’. Universities have responded to this demand by integrating professional communication modules into various degree programmes, often incorporating experiential learning components such as client-based projects and internship requirements. The tangible benefits of this educational approach are evident in both quantitative and qualitative metrics, with graduates reporting higher confidence levels in workplace interactions and employers noting enhanced team collaboration and project management capabilities among newly hired professionals who have completed structured communication coursework. The evolution of professional communication as an academic field has been particularly significant in addressing the technological demands of modern workplaces, with curricula now incorporating elements of digital literacy, virtual team management, and multimodal communication strategies that directly align with contemporary business practices.

Professional communication strategies gain additional theoretical depth when examined through the lens of Levinson’s pragmatic concept of activity types (1992), which provides a valuable framework for understanding how communication patterns and expectations vary across different professional contexts. Activity types, as defined by Levinson (1992), are socially constituted, goal-defined, delimited events with constraints on participants, setting, and

allowable contributions. In professional settings, these manifest as distinct communicative scenarios – from formal board meetings to casual workplace interactions, each with its own set of implicit rules and expectations. Research in professional communication has shown that students who understand these activity-specific constraints and conventions are better equipped to navigate various workplace situations effectively. For instance, recognizing that a job interview represents a specific activity type with its own set of pragmatic rules (turn-taking patterns, appropriate topics, expected levels of formality) helps graduates adapt their communication style appropriately. Similarly, understanding how activity types differ across cultures has become increasingly important in global business contexts, where misalignment of pragmatic expectations can lead to communication breakdowns. Universities that incorporate this pragmatic dimension into their professional communication curricula help students develop a more nuanced understanding of workplace discourse, enabling them to recognize and adapt to different communicative contexts more effectively.

However, at any level, yet another type of context that can provide both pupils and students with the opportunity to discover more about their interests and potential, is represented by academic conferences, where experts in the field of their expertise gather to share the latest updates regarding their research, while passive participants, such as cohorts of candidates to the labour market, can attend, make notes, and derive valuable insights to their own practical benefit.

3. Educational Partnership: Today's Student, Tomorrow's Researcher

Referred to as to an interdisciplinary domain that integrates methods and techniques from linguistics, computer science, and artificial intelligence, with applications in technologies such as machine translation, semantic analysis, and speech recognition, Computational Linguistics represents an interdisciplinary approach to the study of language with technologies derived from computer science which has significantly attracted the attention of expert-researchers on both sides of the fields.

According to the overview offered by Nechifor (2024), Computational Linguistics in Romania has seen significant development since the establishment of the *Informatisation Commission for the Romanian Language* in 2001 at the Romanian Academy. The field was structured in 2002 into four main sectors: Theoretical Linguistics, Linguistic Resources, Language Processing Tools, and Speech Processing, setting the foundation for systematic development in this area. Over the past two decades, Romania has developed substantial linguistic resources and projects. The *CoRoLa* project serves as a comprehensive computational

reference corpus for contemporary Romanian language, while Romanian BERT (*RoBERT*) has been successfully implemented for various applications including sentiment analysis, diacritic restoration, and text generation. The country's participation in major infrastructure projects like *CLARIN* and *DARIAH-Be*, along with the development of lexicographic resources such as *DLR*, *DOOM*, and *CLRE*, has significantly enhanced its Computational Linguistics capabilities.

Still according to Nechifor (2024), the current state of Computational Linguistics in Romania is quite advanced, with Romanian no longer being considered an under-resourced language. The field boasts well-developed recognition systems and has made significant progress in language synthesis. Particularly noteworthy are the developments in speech technology, including sophisticated capabilities in emotion analysis and voice signal processing. Academic infrastructure supporting Computational Linguistics has been robust, with the Master's programme in Computational Linguistics at "Alexandru Ioan Cuza" University running since 2001. Doctoral programmes are offered at several prestigious institutions, including the Technical University of Cluj-Napoca, "Babeş-Bolyai" University, University of Bucharest, and Politehnica University of Bucharest, ensuring a steady path of skilled researchers in the field. Research in Romanian Computational Linguistics has been characterized by strong cooperation between linguists and computer scientists, spanning over two decades. This collaboration has facilitated participation in international consortia and fostered productive partnerships between Romania and Moldova. Notable collaborative projects include *eDTLR*, *DRuKoLa*, *CoRoLa*, *ReTeRom*, *DeLoRo*, *SemRO*, and *CLARIN-RO*.

Despite these achievements, the field faces several challenges. There is a pressing need for a national strategy for Romanian language digitalisation and better coordination between academic institutions. The field requires increased international visibility and improved funding compared to the private sector. Additionally, there is an ongoing challenge to protect the Romanian language against dissolution while simultaneously integrating it into global technological trends. These challenges highlight the need for continued development and strategic planning in Romanian Computational Linguistics.

3.1. Researchers Initiate

This year's edition of the International Conference of the Consortium for the Computerization of the Romanian Language (ConsILR), held in Alba Iulia from November 14 to 16, marked the first edition to include a special section dedicated to high school students. This innovative initiative aimed to familiarize 11th and 12th-grade students with the fascinating field of Computational Linguistics and to explore the practical aspects of research in this interdisciplinary field.

The idea materialized in an educational partnership entitled *Today's Student, Tomorrow's Researcher*, proposed by ConsILR to several high schools in the counties of Iași (Liceul Teoretic "Miron Costin" Pașcani, Colegiul Național "Mihail Sadoveanu" Pașcani, Colegiul Național "Ștefan cel Mare" Hârlău, Liceul Teoretic de Informatică "Grigore Moisil" Iași) and Alba (Colegiul Național "Horea, Cloșca și Crișan" Alba Iulia, Colegiul Teologic Ortodox Alba Iulia). This joint endeavour demonstrated the ability of pre-university educational institutions to collaborate effectively with the academic environment to provide a favourable context for guiding students toward exploring career opportunities in the field of Computational Linguistics.

The activities included in the special section programme for students were designed to provide a comprehensive experience and stimulate pupils' interest in scientific research. The agenda of this section included both a theoretical and a practical session, separated by a break dedicated to informal interaction between pupils and invited experts, who openly discussed the academic and professional journey of researchers in the field of Computational Linguistics. The event concluded with a guided tour of the university campus, during which students had the opportunity to appreciate the modern facilities of the laboratories, meet faculty members, and interact with current students.

In the theoretical session, an introductory presentation on the field of Computational Linguistics was given, with focus on practical applications and the way technology supports the processing of natural language. The session was coordinated by two distinguished guests from the field of Computational Linguistics, who provided students with an in-depth perspective on this innovative domain.

Dan Cristea, a scientific researcher at the Institute of Theoretical Computer Science of the Romanian Academy, Iași branch, presented the theoretical foundations of Natural Language Processing (NLP) through artificial intelligence (AI), emphasizing the contributions of neural networks and the importance of linguistic resources in this field. His speech began by highlighting natural language's 'disordered' beauty, which makes its conversion into algorithmic models difficult. He then explained the difference between *symbolic systems* (rule-based) and *neural systems* (data-driven). While the former explicitly replicates human knowledge, the latter learns linguistic behaviours by processing data. A neural network consists of *successive layers of neurons*: the input layer (which receives the data), the hidden layers (which perform complex calculations), and the output layer (which produces the results). Each neuron functions as a 'mini-computer', adjusting *weights* and *biases* to achieve the most accurate results. The weights determine the importance of each signal, while the biases adjust the activation threshold of the neurons.

Furthermore, it was demonstrated that activation functions, such as *ReLU* or *softmax*, play a crucial role in determining the network's outputs. The *softmax* function standardizes the results within the range $[0,1]$ and transforms them into probabilities, with their sum always equal to 1.

A central aspect of the presentation was the explanation of the learning process in neural networks through the analogy of an exam students took. This process involves three main steps: (1) *forward propagation* (the neural network processes the input data and produces an initial response, similar to students answering questions without knowing all the required information); (2) *error (loss) calculation* (the network compares the obtained result with the correct one and identifies the discrepancies); (3) *backpropagation* (the network adjusts the weights and biases to minimize errors in future iterations). This gradual adjustment method is essential for improving accuracy and refining natural language processing models.

Thus, through relevant comparisons, the fundamental role of linguistic resources and neural networks in natural language processing was highlighted, providing the participants with a deep understanding of how technology can replicate human linguistic behaviours through automated learning processes.

In the second part of the session, Verginica Barbu Mititelu, scientific researcher at “Mihai Drăgănescu” Institute for Artificial Intelligence Research in Bucharest, addressed the role of digital linguistic resources in the development of artificial intelligence systems, offering practical examples and highlighting the contributions of Romanian research in this field. Her presentation brought to the forefront the intersection between linguistics and technology, illustrating how linguistic research contributes to the development of digital resources and applications in the field of artificial intelligence.

The researcher emphasized the importance of involving linguists in interdisciplinary projects, stressing that the development of technology cannot be separated from a profound understanding of language. She outlined the essential steps in creating and utilizing digital linguistic resources, such as electronic dictionaries, lexicons, WordNets, and corpora. These resources are used in projects for automatic natural language processing, contributing to the improvement of applications such as machine translation, speech recognition, and syntactic analysis.

One of the emblematic examples presented was the *RoLEX* project, which analyses the different morphological forms of words in the Romanian language. This resource is used for training algorithms and improving linguistic models used in computer applications. The *Romanian WordNet* and the *CoRoLa* corpus were also discussed, both being fundamental resources in the digital description of the Romanian language. These corpora are built through a complex process that

involves the collection, cleaning, and annotation of linguistic data, either manually or automatically, to ensure the accuracy and consistency of the information.

The presentation also highlighted the challenges encountered in the corpus annotation process, emphasizing the need for consistency in applying labels and ensuring comparability between different resources. Annotation involves the detailed recording of various aspects of the text, such as morphological and syntactic information, which enables precise and efficient language analysis. Additionally, the researcher stressed the importance of establishing usage licenses that ensure the accessibility and transparency of resources for both the academic community and the general public.

Another significant aspect addressed was the necessity of developing natural language processing tools. Verginica Barbu Mititelu explained that linguistic resources have both a descriptive and an applicative value, contributing to the training, testing, and refinement of computer applications. These tools enable the automatic analysis of texts, supporting fields such as machine translation, sentiment analysis, and information synthesis.

In conclusion, the presentation delivered by Verginica Barbu Mititelu during the special session of ConsILR 2024 highlighted the fundamental contribution of linguists to the development of artificial intelligence. Through the projects and digital resources presented, she demonstrated how linguistic research supports technological progress, providing students with a valuable perspective on the potential of the humanities in the digital age. Thus, the meeting between linguists, engineers, and computer scientists becomes a model of interdisciplinary collaboration, capable of transforming the way we understand and use language in the digitised world.

The two theoretical presentations facilitated a clear and accessible understanding of the complexities of this interdisciplinary field, capturing the students' interest and encouraging dialogue on current scientific topics, only to be followed by two applied representations in this field, offered by two researchers from the Institute of Theoretical Informatics and the Faculty of Computer Science in Iași, respectively, Daniela Gîfu and Diana Trandabăț, who coordinated the practical session. Students participated in interactive activities that explored the fundamental tools used in research and simple applications for processing linguistic data.

Thus, the purpose of the practical section was to offer participants the opportunity to experiment with various language processing technologies. The first platform pupils interacted with was *Prodigy Demo*¹. Through its use, they practiced text annotation by marking entities and categories for various purposes.

¹ <https://demo.prodi.gy/>

Using the interactive tool *TensorFlow Playground*, as a second platform², pupils discovered how neural networks function, this representing a practical approach to understanding how artificial intelligence algorithms learn from data. Also, by watching a demonstration of machine translation³, pupils gained insights into the process of automatic translation and the associated challenges, such as accurately interpreting meaning in different cultural contexts. Additionally, by accessing the *Moral Machine Demo*⁴, pupils explored the ethical dilemmas faced by autonomous vehicles, being invited to make ethical decisions in critical scenarios, thereby providing feedback for the development of ethical algorithms.

This conference section represented a model of good practice in the collaboration between pre-university education and the academic environment. The event brought together 60 students from Alba Iulia, who were physically present at the conference, and 350 students from Iași County, who participated online. By exposing students to a cutting-edge field such as Computational Linguistics, the organizers facilitated young people's access to a unique educational experience designed to stimulate their research interest. The initiative also contributed to bridging the gap between pre-university and university education, providing students with a perspective on opportunities for academic and professional development. Furthermore, the partnership represented an important step toward connecting young generations with academic research, inspiring students to further explore the possibilities offered by linguistic technologies and to consider a career in this emerging field.

3.2. Pupils Respond

Against this background, it is interesting to see how young pupils become actively involved in the field, not only as passive and indirect contributors to corpora and data bases compilations under the form of either raw productions of speech patterns or age-appropriate texts that they might be interested in, but also as young researchers, actively involved in the study of language via modern digital means. As mentioned, while there is no evidence of articles in Computational Linguistics authored solely by children, their contributions are evident through the data they provide for research. For instance, the *ChiSCor* corpus comprises 619 fantasy stories told by 442 Dutch children aged 4 to 12, serving as a valuable resource for Computational Linguistics and cognitive science studies (van Dijk et al., 2023). Additionally, the Child Language Data Exchange System (*CHILDES*) offers a repository of child language data that researchers utilize to analyse language

² <https://playground.tensorflow.org/>

³ <https://www.youtube.com/watch?v=GMi4MtyDg40>

⁴ <https://www.moralmachine.net/>

acquisition and development (MacWhinney and Snow, 1990). These contributions are instrumental in advancing the field, even though children are not direct authors of scholarly articles.

However, the active role has already been assumed by young researchers, and thus, under the supervision of professor Livia Dioşan, pupils Stan Albert Alexandru, Gligor Gabriela, and Creţiu Darius, studying in the 10th, 12th, and 11th grades at “Horia, Cloşca şi Crişan” National College, from Alba Iulia, decided to participate in this part of the Conference with a presentation entitled *Computational Linguistics: When Language Becomes Code*. Their contribution was meant as an introduction to the vast domain of Computational Linguistics, under the form of general research based on descriptive data and definitions of the field, with both an informational purpose and a more practical aim, that of opening the field up to their peers, drawing the attention towards it and showing public, manifested interest in it at such a young age.

Their presentation tackled the main perspectives over what Computational Linguistics is, with definitions, main objectives and its positioning at the crossroads among linguistics, artificial intelligence and computer science, at the same time offering a short history of this growing domain. According to their perspective, Computational Linguistics has evolved significantly over the decades, beginning with early research in the 1950s-1960s on automatic translation and syntactic analysis. By the 1970s-1980s, advanced algorithms enabled text and natural data processing, and from the 1990s onward, statistical models and deep learning revolutionized the field. Key components include syntax (sentence structures), semantics (contextual meaning), and pragmatics (contextual interpretation), supported by techniques such as natural language processing (NLP) and neural networks. Its applications are vast, ranging from virtual assistants (e.g., Siri, Alexa) and automatic translation tools to sentiment analysis, speech recognition, and process automation. However, challenges persist, such as language ambiguity, linguistic diversity, data quality, and understanding nuanced contexts. Future advancements will focus on multimodal models, improved applications, new domains like healthcare and education, and addressing ethical concerns like privacy and bias. Ultimately, Computational Linguistics transforms human language into a machine-readable code, already prevalent in technology today, with even greater potential for enhancing human-computer interaction in the future.

What is notable about Stan, Gligor, and Creţiu’s presentation is the interest manifested in the increasing relevance of Computational Linguistics in today’s rapidly advancing technological landscape. This active engagement in cutting-edge topics such as natural language processing, machine learning, and neural networks is invaluable for preparing pupils to enter the labour market. As

the demand for expertise in areas like virtual assistants, automated translation, and AI-driven solutions grows, the pupils are positioning themselves at the forefront of a field with vast career opportunities. Their work exemplifies how education can bridge the gap between academic inquiry and practical industry needs, educating a generation ready to tackle both technological advancements and ethical considerations in the workforce.

Manifesting an even bolder approach, 12th grader pupils Ecaterina Ștefănescu and Alexandru-Iulius Jerpelea from “Tudor Vianu” National College of Computer Science, in Bucharest, successfully attempted a participation in ConsILR 2024 with an academic paper, under the ‘Applications of NLP and AI in Various Domains’ Session, on the second day of the event, chaired by Tudor Bumbu, delivering a presentation on and defending their ideas regarding: *Reddit is All You Need: Authorship Profiling for Romanian*. Their study was based on the assumption that authorship profiling involves determining an author’s characteristics based on their written content, so they built a novel corpus of short texts in Romanian, annotated with specific author characteristic keywords for which they used the social media platform Reddit, using to their advantage its community-based thematic structure (subreddits) to gather information about authors’ backgrounds. From these cues, they inferred users’ demographics and broader personal traits, such as age category, employment status, interests, and social orientation, the process resulting in a dataset comprising over 23,000 samples drawn from more than 100 Romanian subreddits, which they analysed and fine-tuned Large Language Models (LLMs) to establish baseline performance for authorship profiling, demonstrating the corpus’s potential for advancing research in this field.

Having all their resources publicly available, they decided to analyse the text from the perspective of the categories: subdialect of the Daco-Romanian dialect, educational or employment status, personal interests, such as computer science and music, and personal inclinations, such as social orientation and technological interest. To create a robust dataset for authorship profiling in Romanian, the researchers began by identifying relevant subreddits. A total of 263 Romanian subreddits were initially identified through manual searches and a subreddit exploration tool, which were then filtered based on relevance, post quality, and volume, resulting in 146 selected subreddits. These subreddits were categorized to reflect various author characteristics. For instance, subdialects were identified based on regional linguistic registers associated with four main Romanian dialects: Transylvanian, Moldavian, Oltenian, and Wallachian. Additionally, subreddits were tagged to indicate users’ education or employment status (e.g., pre-university students, university students, employed individuals) and general interests, represented by 32 specific labels grouped into six broader fields of

interest (e.g., science, economy). To capture personal inclinations, 12 categories, such as an active lifestyle or an interest in technology, were assigned based on subreddit characteristics. Posts were then extracted using the Reddit API, limiting up to 1,000 posts per subreddit and including only those with a minimum of 75 characters to ensure relevance. The dataset was cleaned by removing non-Romanian content, URLs, stray symbols, and other irrelevant elements, resulting in a final dataset of 26,517 posts. Informal language mixing was tolerated to preserve linguistic information. Challenges in assigning specific keywords, such as regional dialects, led to a simplified structure where all keywords were merged. This approach shifted the authorship profiling task from classification to feature extraction, making it more suitable for analysis using Large Language Models (LLMs).

A baseline authorship profiling tool was developed to extract keywords that reflect an author's characteristics based on their text, generating these keywords only when relevant. The Large Language Model (LLM) Llama 3.1-8B-Instruct was fine-tuned using the Low-Rank Adaptation (LoRA) technique, which allowed for computational efficiency by updating only a small subset of parameters. The training process utilized a dataset split into training, evaluation, and testing subsets (95:5:5 ratio), with training stopped after one epoch when evaluation loss plateaued. Posts longer than 2,200 tokens were excluded due to memory constraints. The model's performance was evaluated using exact match accuracy (78.5%) and Intersection over Union (IoU, 80.4%), which accounted for overlap between predicted and expected keywords. While the model generated 52 unique keywords aligning with the corpus, it did not create entirely new keywords and tended to focus more on content-related rather than author-focused keywords, a pattern reflecting the dataset's annotations. Limitations included the inability to combine multiple traits, such as subdialect and interests, in the same prediction, often due to gaps in annotation. Despite these challenges, the tool demonstrated promising results, indicating significant potential for advancing Romanian authorship profiling with refined datasets and modelling approaches.

Ștefănescu and Jerpelea's research presented the first Romanian-language corpus for authorship profiling, annotated with keywords from various categories, and introduced a baseline Large Language Model (LLM) trained for keyword generation as a feature extraction task. According to the authors' own observations, despite its promising potential, the study revealed the need for improved data diversity, more robust annotation methods, and exploration of alternative model architectures to overcome limitations such as annotation bias, classification ambiguity, and computational constraints, but the originality of this work is notable, particularly as a pupil-led analysis, showcasing an impressive effort in tackling a complex problem like authorship profiling while contributing

a novel resource to the field. However, ethical concerns around potential biases in manual tagging highlight the importance of careful refinement in future research, but, nevertheless, encourage peer pupils to develop skills in this field, inviting to further research at a young level.

4. Conclusions

School and career orientation represents an essential process in preparing students for the complex decisions regarding education and career. It integrates self-awareness and exploration of options to ensure a well-grounded path. A crucial role is played by students' interaction with experts in the field that interests them. This study explored the multifaceted nature of work and its significance in shaping individual identity. It emphasized the importance of early career guidance, particularly through school and vocational orientation programmes, to equip young people with the necessary knowledge and skills for successful career transitions. The study highlighted the crucial role of collaborative efforts between educational institutions, industry professionals, and researchers in providing students with meaningful experiences, such as internships, industry visits, and academic conferences, that bridge the gap between theoretical learning and real-world applications.

An important acknowledgement referred to the indirect but significant contributions of young people to the field of Computational Linguistics through the provision of valuable linguistic data. While not directly involved in academic research, pupils and students play a crucial role in generating the raw material that fuels advancements in this field, thus underscoring the importance of cultivating a supportive and collaborative environment that encourages young people to explore their interests, develop their skills, and make informed decisions about their future careers. But, in order to overcome these passive roles, providing them with access to diverse learning experiences and engaging them in meaningful research activities, they can be empowered to become active participants in shaping their own futures and contributing to the advancement of knowledge in various fields.

Last, but not least, the article examined the evolving field of Computational Linguistics, emphasizing the interdisciplinary nature of this domain and its growing importance in the digital age. By showcasing the successful collaboration between ConsILR and high schools in the *Today's Student, Tomorrow's Researcher* initiative, the potential of such partnerships to inspire young people to pursue careers in research and contribute to advancements in this field was under investigation.

References

- Brown, S. D., and Lent, R. W. (2013). *Career Development and Counselling: Putting Theory and Research to Work*. Wiley.
- Gati, I., and Asher, I. (2001). The PIC model for career decision-making: Pre-screening, in-depth exploration, and choice. *Journal of Counselling Psychology*, 48 (4), 490-505.
- Levinson, S. C. (1992). Activity types and language. In P. Drew and J. Heritage (Eds.). *Talk at work: Interaction in institutional settings*. Cambridge University Press, pp. 66-100
- National Association of Colleges and Employers. (2022). Job Outlook 2022: The attributes employers want to see on students' resumes. *NACE Journal*: 82(3), 23-32.
- Nechifor A., C. Dimulescu (2024). External Factors Influencing First-Year Students' Usage of Digital Learning Tools at 'Transilvania' University of Braşov. In Blažević I., M. Diković, I. Kovačić, L. Lazarić, and J. Gugić (eds.). *Proceedings of International Scientific Conference 20th Mate Demarin Day 'Diversity and Interdisciplinarity in Education*. Pula, Croatia: Juraj Dobrila University in Pula Publishing House, pp. 251-274
- MacWhinney, B., and Snow, C. (1990). The Child Language Data Exchange System: An update. *Journal of Child Language*, 17(2), 457-472.
- Munteanu, D., A. Nechifor. (2024). The Entrepreneurship vs. Intrapreneurship? How do Today's Students Perceive the Two Concepts in Terms of Access in the Labour Market. Presentation at *The 12th European Conference on Education*, 11-15 July, 2024, London, UK
- Munteanu, D., A. Nechifor (2023). Continuous Improvement of Processes and Activities within 'Transilvania' University of Braşov Based on Student Ideas. Be Central! Internal Project Competition. *EDULEARN23 Proceedings*. p. 29-38
- Munteanu D., A. Nechifor (2019). Motivational Projects for Students at 'Transilvania' University of Braşov, Romania. *EDULEARN19 Proceedings of the 11th International Conference on Education and New Learning Technologies*. p. 201-209
- Purcaru, M.A.P., A. Nechifor (2013). Rethinking Career Paths: The Methodology of Teaching Mathematics and English to Adult Teachers through Conversion Programmes - A Comparative Study. *Bulletin of the 'Transilvania' University of Braşov. Series VII: Social Sciences and Law* 6 (55) No. 2: 27-34
- Rahman, S., Chen, K., and Thompson, M. (2023). Impact of professional communication training on graduate employability: A longitudinal study. *Journal of Business and Technical Communication*: 37(2), 145-168.
- Savickas, M. L. (2013). Career Construction Theory and Practice. *Career Development and Counseling: Putting Theory and Research to Work*. Wiley.
- Super, D. E. (1980). A Life-Span, Life-Space Approach to Career Development. *Journal of Vocational Behavior*, 16(3), 282-298.
- van Dijk, B. M. A., van Duijn, M. J., Verberne, S., and Spruit, M. R. (2023). ChiSCor: A Corpus of Freely Told Fantasy Stories by Dutch Children for Computational Linguistics and Cognitive Science. arXiv preprint arXiv:2310.20328.
- Watts, A. G., and Sultana, R. G. (2004). Career Guidance Policies in 37 Countries: Contrasts and Common Themes. *International Journal for Educational and Vocational Guidance*, 4(2-3), 105-122.
- Watson T. J. (2003). Sociology, Work and Industry. In Taylor and Francis, *e-Library*.
- Zamfir, C., Vlăsceanu, L. (1998). *Dicţionar de sociologie*, Editura Babel, Bucureşti.

*** (2016). Ordinul Ministrului Educației Naționale și Cercetării Științifice nr. 5.034/29.08.2016 pentru aprobarea Metodologiei de organizare a Programului național “Școala altfel”. Source: <https://www.edu.ro/etichete/%C5%9Fcoala-altfel>

INDEX OF AUTHORS

AMARIE, RADU-SEBASTIAN	171	LUCA, RAMONA	23
BADEA, ADRIAN-CĂTĂLIN	171	MALAHOV, LUDMILA	7, 23
BEJINARIU, SILVIU-IOAN	23	MORUZ, ALEX	64
BELCIUG, VLAD	138	MUSCALU, ȘTEFAN-GABRIEL	171
BOLEA, CECILIA	203	NAGACEVSCI, VALENTIN	154
BUCUR, ANA-MARIA	107	NECHIFOR, ANDREEA	221
BUMBU, TUDOR	7, 23	NIȚĂ, SARA	45
BURTSEVA, LIUDMILA	7	OLARIU, FLORIN-TEODOR	23
CERESCU, MARIUS	7	PĂIȘ, VASILE	45
CHITEZ, MĂDĂLINA	107	PANĂ, LUCA	45
CINKOVÁ, SILVIE	95	PELICAN, ELENA	138
COJOCARU, SVETLANA	7	PLECHÁČ, PETR	95
COLESNICOV, ALEXANDRU	7	ROGOBETE, ROXANA	107
CRISTEA, DAN	64, 203	SCUTELNICU, ANDREI	203
CUREA, ERIC	45	SHRIPAD, ASMII	188
DĂSCĂLESCU, ȘTEFAN-COSMIN	171	ȘOITU, LAURENȚIU	64
DINCĂ, ANDREEA	107	ȘTEFĂNESCU, ECATERINA	125
DINU, ALEXANDRU	79	TOMA, TUDOR P.	154
DUMITRAN, ADRIAN MARIUS	171	UNGUREANU, MARIA	64, 221
DUMITRAN, ANGELA-LILIANA	171		
FRÎNCU, IOANA-SIMINA	107		
GÎFU, DANIELA	203		
JERPELEA, ALEXANDRU-IULIUS	45, 125		