

Editors:

Anca Vasilescu

Daniela Gîfu

Dan Tufiş

Dan Cristea

LINGUISTIC RESOURCES AND TOOLS FOR NATURAL LANGUAGE PROCESSING

Proceedings of the 18th International Conference,
Braşov, 11-14 December 2023



Editura Universităţii „Alexandru Ioan Cuza” din Iaşi
2024

Organisers:

Faculty of Mathematics and Computer Science
Transilvania University of Braşov

Institute for Computer Science
Romanian Academy, Iaşi branch

Faculty of Computer Science
“Alexandru Ioan Cuza” University of Iaşi

Research Institute for Artificial Intelligence “Mihai Drăgănescu”
Romanian Academy, Bucharest

Romanian Association for Computational Linguistics

Academy of Technical Sciences of Romania

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 Iaşi

ISSN: 3045-2074

ISSN-L: 3045-2074

PROGRAM COMMITTEE

Vasile Apopei, Institute for Computer Science, Romanian Academy, Iași

Verginica Barbu Mititelu, Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Bucharest

Victoria Bobicev, Technical University of Moldova

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

Svetlana Cojocar, Academy of Sciences of Moldova & Institute of Mathematics and Computer Science “Vladimir Andrunachievici”, Chișinău

Dan Cristea, Institute for Computer Science, Romanian Academy, Iași & Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

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

Daniela Gifu, Institute for Computer Science, Romanian Academy, Iași Branch

Gabriela Groseck, West University of Timișoara

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

Diana Zaiu Inkpen, School of Electrical Engineering and Computer Science, University of Ottawa
Radu Ion, Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Bucharest

Elena Irimia, Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Bucharest

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

Constantin Orășan, University of Surrey

Elena Pelican, Ovidius University of Constanța

Vasile Florian Păiș, Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Bucharest

Mircea Petic, “Vladimir Andrunachievici” Institute of Mathematics and Computer Science, Moldova State University, Chișinău, Republic of Moldova

Andrei Scutelnicu, Institute for Computer Science, Romanian Academy, Iași & Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

Elena Isabelle Tamba, “Alexandru Philippide” Institute for Romanian Philology, Romanian Academy, Iași

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

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

Ștefan Trăușan-Matu, Faculty of Automation, Control and Computer Engineering, Politehnica University of Bucharest & Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy

Dan Tufiș, Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Bucharest

Anca Vasilescu, Faculty of Mathematics and Computer Science, Transilvania University of Brașov

ORGANISING COMMITTEE

Vasile Apopei, Institute for Computer Science, Romanian Academy, Iași

Liana Luminița Boca, Faculty of Mathematics and Computer Science, Transilvania University of Brașov

Cecilia Bolea, Institute for Computer Science, Romanian Academy, Iași

Dan Cristea, Institute for Computer Science, Romanian Academy, Iași & Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

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

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

Daniela Gifu, Institute for Computer Science, Romanian Academy, Iași Branch

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

Andreea Nechifor, Faculty of Letters, Transilvania University of Brașov

Maria Nuțu, Faculty of Mathematics and Computer Science, Transilvania University of Brașov

Andrei Scutelnicu, Institute for Computer Science, Romanian Academy, Iași & Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

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

Dan Tufiș, Research Institute for Artificial Intelligence “Mihai Drăgănescu”, Romanian Academy, Bucharest

Anca Vasilescu, Faculty of Mathematics and Computer Science, Transilvania University of Brașov

TABLE OF CONTENTS

FOREWORD	VII
CONSILR MEETING – AN OVERVIEW. MAPPING THE FUTURE: DIGITAL HUMANITIES STRATEGY FOR ROMANIA.....	1
<i>Resumed by Andreea Nechifor</i>	
KEYNOTE SPEECHES	15
CHAPTER 1. DIGITALISATION OF CULTURES	21
DIGITALISATION OF CULTURES. PREPARATIONS FOR THE ADOPTION OF A STRATEGY FOR THE ROMANIAN LANGUAGE	23
<i>Dan Cristea, Gabriela Haja, Elena Isabelle Tamba, Dan Tuفیş</i>	
CHAPTER 2. NATURAL LANGUAGE PROCESSING FOR E-LEARNING	41
EDUCATIONAL TEXT CONTENT GENERATION FOR ELEARNING SYSTEMS	43
<i>Alexandr Parahonco, Mircea Petic</i>	
METAPHOR MASTERY: AUGMENTED REALITY DICTIONARY FLASHCARDS FOR E-LEARNING	63
<i>Oleseca Caftanatot, Ludmila Malahov, Inga Țițchiev, Dan Tălămbuță, Daniela Caganovschi, Eugenia Mincu</i>	
ENHANCING LANGUAGE LEARNING THROUGH AUGMENTED REALITY: A FOCUS ON MULTIWORD EXPRESSIONS	73
<i>Oleseca Caftanatot, Ludmila Malahov, Inga Țițchiev, Dan Tălămbuță, Tudor Bumbu, Ana Nastasiu, Victoria Bobicev</i>	
CHAPTER 3. ANONYMISATION AND SUMMARISATION	83
TOWARDS ANONYMISING THE ROMANIAN JURISPRUDENCE	85
<i>Vasile Păiș, Verginica Barbu Mititelu, Elena Irimia, Radu Ion, Valentin Badea, Dan Tuفیş, Cosmin Sterea-Grossu</i>	
COMPARISON OF MANUAL AND AUTOMATIC SUMMARISATION	101
<i>Victoria Alexei, Victoria Bobicev</i>	
CHAPTER 4. LANGUAGE MODELS	117
THE NEWS IN BRIEF – LEVERAGING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE IN NEWS CLUSTERING, SUMMARISATION AND EVALUATION	119
<i>Bogdan Florea, Adrian Iftene</i>	

MYANMAR SIGN LANGUAGE RECOGNITION USING DEEP LEARNING	135
<i>Bhone Myint Swe, Min Thaw Phyto, Min Chan Lun, Aung Sann Thit, Aung Hlaing Moe, Kaung Khant Min, Phyto Thu Htet, Aye Su Phyto</i>	
A ROMANIAN BERT MODEL FOR LINGUISTIC ANALYSIS	149
<i>Radu Ion</i>	
CHAPTER 5. DISCOURSE AND HUMOUR	163
ANNOTATING DISCOURSE RELATIONS IN THE ROMANIAN REFERENCE TREEBANK.....	165
<i>Tudor Voicu, Verginica Barbu Mititelu</i>	
THIS MAKES ME LAUGH. DOES IT MAKE YOU LAUGH, TOO?	177
<i>Daniela Gîfu, Dan S. Stoica</i>	
DETECTION OF MEANING RELATIONS IN ROMANIAN WORD DEFINITIONS	187
<i>Călina-Georgiana Ciocoiu</i>	
CHAPTER 6. ROMANIAN SPEECH.....	197
MAKING LINGUISTIC RESOURCES ACCESSIBLE. THE AUDIO ARCHIVE OF THE NEW ROMANIAN LINGUISTIC ATLAS BY REGIONS. MOLDOVA AND BUKOVINA	199
<i>Florin-Teodor Olariu, Alexandru Laurențiu Cohal, Luminița Botoșineanu, Veronica Olariu, Ramona Luca, Silviu-Ioan Bejinariu</i>	
TOWARDS SPEECH-BASED WEB APP DEVELOPMENT	215
<i>Irina Vasilița, Adrian Iftene</i>	
INVESTIGATION OF ROMANIAN SPEECH RECOGNITION IMPROVEMENT BY INCORPORATING ITALIAN SPEECH DATA	235
<i>Carol-Luca Gasan, Vasile Păiș</i>	
INDEX OF AUTHORS.....	249

FOREWORD

The volume includes the papers of the 18th edition of the conference "Linguistic resources and tools for natural language processing", which this year took place in Braşov, a wonderful city located in the center of Romania. From its inception, in 2001, the ConsILR Conference (the acronym for *Consortiul de Informatizare pentru Limba Română*) an initiative born in the Section for Information Science and Technology of the Romanian Academy) was meant as a meeting place for linguists and computational linguists, but also for researchers of the humanities, PhD students and master students in Computational Linguistics, all with interest in the study of the Romanian language from a computational perspective. The ConsILR series has run, with few exceptions, every year, first in the format of a workshop, and since 2010 as an international conference. ConsILR is not strictly addressed to researchers working on Romanian language but also to other scientists, from any part of the world, who share their experiences in modelling and implementing techniques for their own languages that might inspire scientists working on various other languages. Today, many modern language technologies have a universalist perspective, with generic solutions applicable to as many languages as possible (see for instance the Universal Dependencies project, or the recent COST action UniDive, accounting for and fostering language diversity).

The host organiser was Transilvania University of Braşov with the Department of Mathematics and Computer Science, who worked in close collaboration with colleagues from the Faculty of Computer Science of the "A.I. Cuza" University in Iaşi, the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy, Institute of Computer Science (Iaşi Branch of the Romanian Academy), "A. Philippide" Institute of Romanian Philology (Iaşi Branch of the Romanian Academy), and the Romanian Association of Computational Linguistics. An important contribution was made by the members of the scientific committee, thanks to whose efforts the quality of the papers was substantially improved, and we take this opportunity to thank them for their assistance.

The first day of the Conference was dedicated to an extended session of the Consortium for Informatisation of the Romanian Language, which is resumed at the beginning of the volume.

The conference program included 15 papers, grouped into 6 sections, in this order:

- **Digitalisation of Cultures,**
- **Natural Language Processing for e-Learning,**
- **Anonymisation and Summarisation,**

- **Language Models,**
- **Discourse and Humour** and
- **Romanian Speech.**

Also, five keynote speeches were presented during the conference, namely:

- Gabriela Haja – **Stages of Digitization in Romanian Academic Lexicography,**
- Marius Popescu – **AD-NLP: A Benchmark for Anomaly Detection in Natural Language Processing,**
- Radu Ion – **The Romanian Language in the Age of Standard and Large Language Models,**
- Liviu Dinu – **Computational Approaches in Historical Linguistics,**
- Radu Tudor Ionescu – **Recent Text and Audio Resources for the Romanian Language.**

The recent advances of language technology based on large and very large language models could not be overlooked and both the keynote speeches and several regular papers reported on relevant aspects of this mainstream research.

The current selection of papers is inherently incomplete, and its coverage is far from being ideal. There are many more significant results in the NLP area dedicated to Romanian language, but some of them were already published elsewhere. There are also several other reputed computational linguists, who worked on Romanian or with Romanians, who, unfortunately, are not present in this volume, but we acknowledge their work and bring here our homage to them.

To conclude, this volume offers a valuable collection of contributions presenting research and development results. Several issues discussed throughout the book are present-day concerns of the scientific community, irrespective of the investigated language. Therefore, 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.

The papers presented at the conference brought new and valuable contributions to the development of the field of natural language processing, many of them being focused on applications for the Romanian language.

May 2024
The Editors

CONSILR MEETING – AN OVERVIEW

MAPPING THE FUTURE: DIGITAL HUMANITIES STRATEGY FOR ROMANIA

Minutes of the meeting of the Informatization Consortium for the Romanian Language – the first day of the 18th International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2023

Resumed by Andreea Nechifor
Transilvania University of Braşov / Faculty of Letters

Serving as a pivotal bridge between human language and machine understanding, Natural Language Processing (NLP) stands at the forefront of forward-looking research in the fields of artificial intelligence and computational linguistics. Over the past few decades, NLP has evolved from a niche area of study to a dynamic and interdisciplinary field that interconnects linguistics, computer science, and cognitive psychology. Setting as a primary goal the enabling of machines to comprehend, interpret, and generate human language, NLP unlocked a multitude of applications ranging from automated translation and sentiment analysis to chatbots and voice recognition systems.

The setting of NLP dates back to the 1950s when forerunners like Alan Turing (1950) and Noam Chomsky (1956) underpinned the common basis of language and computation. However, it was only in recent years, with the surge in computational power and the introduction of deep learning, that NLP witnessed unprecedented progress. Ultramodern models, such as transformer-based architectures like BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) and GPT (Generative Pre-trained Transformer) (Radford et al., 2018), have revolutionised the field, pushing the boundaries of language understanding, processing and generation.

Contemporary breakthroughs are yet to be praised to their full potential, as challenges are still to be met, but with the emphasis laid on the role of deep learning, recurrent neural networks (RNNs), and with attention dedicated to the mechanisms of enhancing the capabilities of NLP systems, to the nuances of language semantics, to context ambiguity, and to the need for large-scale

annotated datasets for training robust models, drawbacks will successfully be overpassed.

Thus, navigating the intricacies of NLP advancements, a group of researchers establishing the Informatisation Commission for the Romanian Language, in March 2001, gathered for the first time at the headquarters of the Romanian Academy, on Calea Victoriei (Forăscu et al., 2006). And, as it has become obvious along the time that the interplay between linguistic theories and state-of-the-art technologies is shaping the future landscape of language-oriented AI applications, this meeting has grown into a tradition through the annual organisation of workshops and, later, conferences in Bucharest, Iași, and Chișinău by an extended group, which, for this reason, was named Consortium (Forăscu et al., 2006). Their approach was immediately afterwards four-folded, as in 2002 4 sectors have been set-up: Theoretical Linguistics, Linguistic Resources, Language Processing Tools, and Speech Processing (from the interventions of Dan Tufiş & Dan Cristea, in the meeting) and has been intended to contribute to the ongoing dialogue within the NLP community, shedding light on the achievements, challenges, and future prospects that define this dynamic field.

In 2023, the Consortium's gathering took place for the first time in Braşov, on the first day of the 18th International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR-2023). In this edition, the joint parallel event was dedicated to a revitalisation of the Consortium, with one full day of on-site and on-line interventions. Around 30 people had the opportunity to give 5 to 20-minute speeches in the morning and after-noon sessions. Among them, five members of the Romanian and Moldavian academies, including two vice-presidents of these institutions, were invited to share their insights over the status of Digital Humanities and Natural Language Processing in the two countries. Directors of research institutes affiliated to the Academy and of some Romanian and Bessarabian universities, researchers working in linguistics and computer science research institutes from Braşov, Bucharest, Iași, Chișinău, Cluj-Napoca, Bruxelles, Timișoara, Bălți, and Fairfax showed the vivid interest that this field of research has gained in the last years, the maturity of the linguistic resources involved in the study and informatisation of the Romanian language and the high necessity to shake hands more than ever for cooperation and common initiatives. The meeting, hosted by Transilvania University of Braşov and coordinated by Anca Vasilescu, from the Faculty of Mathematics and Computer Science, underscored the belief that aligning linguistics and informatics, both scientifically and organizationally, represents a significant advancement in the field of Digital Humanities (from the intervention of Dan Tufiş in the meeting).

Appreciating the initiative to engage the university community in this field and to foster collaboration in the same domain with counterparts from the Republic of

Moldova, whose language was officially recognized as Romanian in 2023, Nicolae-Victor Zamfir, vice-president of the Romanian Academy, endorsed the expressed interest in this research area. Svetlana Cojocaru, vice-president of the Academy of Sciences of the Republic of Moldova, shared a similar viewpoint, highlighting the benefits of converging and aligning mutual interests. She capitalised on the positive outcomes of a summer school organised in Bălți, in September 2023, and of a previous edition of the conference held in Chişinău, in November 2022. Additionally, Dan Tufiş, president of the Informatisation Commission for the Romanian Language, mentioned that a proposal has been put forth to reconfigure this commission through consultations. At the same time, Dan Cristea, on behalf of the Institute of Computer Science of the Romanian Academy from Iaşi, addressed a plea for a unified vision of the two academies, in Romania and the Republic of Moldova, of Romanian and Moldovan research institutes and universities, of computer scientists and humanists concerned with the Romanian language, with the aim of adopting a unified vision that promotes digital methods applied in the study of Romanian cultural creations and language. According to Dan Tufiş, this involves the formation of a joint commission comprising the academies from both Romania and the Republic of Moldova. As the initiatives progress, there is a call for feedback and proposals to ensure a collaborative and inclusive approach in view of advancing Romanian language technology.

The benefits of the collaboration between linguists and computer scientists are obvious and the organisers of the event can be named visionaries, as this field started to bloom in Romania more than 20 years ago and has successfully developed meanwhile, said academician Nicolae-Victor Zamfir. The linguist-computer scientist nexus, forged through a commitment spanning over two decades, as also recalled by Dan Cristea, has been marked in Romania by joint projects participation in international consortia, and scientific events organised and conducted collaboratively by specialists from both disciplines. Such examples, as mentioned by Cristea, encompass projects like: eDTLR, DRuKoLa, CoRoLa, ReTeRom, DeLoRo, SemRO, E-READ, MappingBooks, CLARIN, ENeL, ACCURAT-RO, STAR, CLARIN-RO, SIR-RESDEC, and METANET4U.

Placed at the intersection of humanities and information technologies, the field of Digital Humanities is, in Dan Cristea's perspective, in continuous and expansive development, including dedicated computer applications for the preservation and exploitation of cultural heritage, providing humanists with sophisticated research tools capable of opening new paths of exploration. The objects listed by Cristea for study in this field may include textual and spoken works, 3D architectural reliefs and digital reconstructions, musical scores, scanned pages from heritage books, virtual museums, etc.

Born with the intention to preserve Romanian written and spoken culture and to foster the harmonious development of the Romanian language, to enable its protection against the threat of dissolution into widely spoken languages worldwide, while simultaneously working towards its integration into long-term global technological trends, as well as to ensure its immunity to influences that may arise from passing trends, it is shown that the idea of digitising a culture was met with resistance, either out of being catalogued as violent, restrictive, perverse, deviant, or as a result of being reduced to a mere tool (from Dan Cristea's speech). However, after clearly delineating the difference between two key concepts in this field: *digitisation* and *digitalisation*, the advantages of submitting one's language to a process of digitalisation become obvious. Thus, understanding *digitisation* as the process of transforming the medium of a cultural artefact into a digital format (digitally photographing a painting, scanning a book, creating a 3D copy of a sculpture) and *digitalisation* as a paradigm shift, a change in perspective, regarding the manipulation of a cultural work through the intelligent interpretation of the work's content, which encompasses digitisation only as one of its basic processes, the benefits of digitalisation can be fully comprehended.

Cristea also mentioned the most important institutional projects conducted within entities with a humanistic profile, among which: *The Electronic Dictionary of the Romanian Language - Studies on Acquisition, Linguistic Resources in Electronic Format. Monumenta linguae Dacoromanorum. Biblia 1688, The Essential Romanian Lexicographic Corpus*, and *The Dictionary of the Romanian Language*, alongside referring to a number of events with joint organisation: the ConsILR conference-series; the EUROLAN schools-series; the edition-series on Romanian academic lexicography and the challenges of computerisation; the meeting-series of the members of the German-Romanian joint project DRuKoLa; the workshop-series organised as satellite events of the ConsILR conferences; the Iași - Chișinău collaborations involving linguists and computer scientists from both countries; the calls for collaboration and promotion for the eDTLR project; the "Solomon Marcus" symposium-series organised by the Faculty of Mathematics and Computer Science in Bucharest; and the SpeD conference-series on speech processing and man-computer dialogue, organised by the National University of Science and Technology Politehnica Bucharest. Following this, Cristea noted that Romania has been actively engaged in an extensive digitisation process of philological research, in perfect alignment with the prevailing European linguistic policy.

Elena Isabelle Tamba from the Institute of Romanian Philology "Alexandru Philippide" in Iași pointed out that in 2019 the Romanian Academy joined with an active status the *European Federation of National Institutions for Language*.

According to its official site¹, EFNIL is an entity that serves as an advisory body of the European Commission for European language policies, with the main objective to promote European linguistic diversity as a means of preserving and extending the richness of European culture and developing a sense of shared European identity. This membership has fostered, as Tamba mentioned, the development of European projects, with Iași and Bucharest actively participating in these initiatives, such as the *European Language Equality* (ELE).

In this respect, Vasile Păiș from the Research Institute for Artificial Intelligence of the Romanian Academy in Bucharest detailed that ELE performs the automatic and manual gathering of resources and technologies for each official language of the EU (including Romanian), subsequently stored in the *European Language Grid* (ELG). Păiș added that, according to the digital language equality dashboard, created in ELE and continuously updated since then, filtered interrogations can be performed also for Romanian with good results for text and speech processing and natural language generation, and with a still lower ranking for human-computer interactions with audio or video input.

With the goal of enhancing accessibility to resources generated for the Romanian language over the past two decades, Tamba addressed the shifts in the contemporary landscape of Romanian lexicography. She emphasised the role of digitalisation as a component of the broader cultural strategy, with the Romanian language user embodying different personae: the general public, researchers, translators or persons involved in the educational process. Hence, funded by public resources, the overarching aim of these initiatives to digitise research is to guarantee unrestricted access to the public at large, regardless of specialisation. And even if, currently, the digital resources created in the specialised institutes of the Romanian Academy can be used freely (in open access), for the moment they are not yet accessible as linked open data. Tamba offered an overview of the research projects for creating various types of resources and digital tools, which have been developed for natural language processing, with application to the Romanian language, such as: digital dictionaries / lexicons (general or specific – DLR, DOOM, etc.), lexicographic corpus (CLRE) and laid stress on the idea that, in the past, the role of the lexicographers concluded once the manuscript was handed over, but today, it extends to educating dictionary users, by offering usage guides, dictionary networks, and lexical resources to them, as part of a new approach in linguistic consultancy.

In the same line of thought, Gabriela Haja, manager of the Institute of Romanian Philology “Alexandru Philippide” from Iași, tried to provide professional solutions for the appetite of the public to generate and use digital

¹ <https://efnil.org/documents/mission/>

dictionaries. Thus, adequate answers to private initiatives like: Dexonline, Wikipedia, Wikisource, Wiki communities and Meta-Wiki, can be the implementation of national and international strategies for the preservation and promotion of languages through digital means, the design of a national strategic plan for the digitisation of the Romanian language, the development of and the investment in resources and (digital) tools, gold corpora, the creation of dictionaries by lexicographers, endorsed by academic institutions, the involvement of specialists (linguistics, NLP, PR), and more funding.

And even though Haja identified as weaknesses the insufficient dialogue among experts in these fields (including at the decision-making level), the existence of isolated projects which do not belong to a strategic digitisation plan, the insufficient infrastructure and personnel for the maintenance, optimisation, and promotion of the results achieved, the insufficient funding compared to the private sector, the perception of digitisation as either a mere tool for popularising knowledge or as a threat, as well as the incidence of damaging internal competition, she was also able to pinpoint the strengths, which are not only worth mentioning, but also encouraging. Thus, the Romanian Academy has competent specialists in each of the digital humanities fields (linguistics, NLP, AI) and its projects are unique in the country. In her opinion, there are visible results available to the public in open access, stable funding, constructive internal competitions and great experience already amassed in collaborative projects, and, in this way, the perspective over digitisation can be steered towards viewing it as a research object or as a means of preserving a language and a culture.

To further support the positive trend towards the introduction of digital approaches in the study the Romanian language, Gabriela Haja staged the evolution of the work done so far in the field of lexicology, performed mainly by the research teams belonging to the Institute of Romanian Philology “Alexandru Philippide” of the Iași branch of the Romanian Academy, in collaboration not only with specialists from the Faculty of Computer Science at “Alexandru Ioan Cuza” University of Iași, but also with national and international teams from the same field. A selection from the listing includes: DLRI, eDTLR, CLRE, DLRi, the SoLiRom Platform of the Romanian Academy, DLR, DA/DLR, DELR, DOOM3, and last, but not least, CLRE.

At the other end of the spectrum, the *LiRo Benchmark for Romanian Language Tasks* is a platform² currently in use, primarily targeting computer scientists rather than philologists. According to Marius Popescu and Radu Ionescu, from the Faculty of Mathematics and Computer Science, University of Bucharest, this is an open-source benchmark and a continuous-submission leaderboard, storing

² <https://lirobenchmark.github.io>

public Romanian language datasets dedicated to specific tasks and configured to represent a repository and consultancy platform to aggregate the existing Romanian language resources. And, as a good way to encourage development is through public leaderboards, LiRo welcomes new tasks, datasets and baselines, with help provided by native Romanian volunteers.

Projects like CoRoLa, CLRE, DeLoRo, the treebanks developed in Iași and Bucharest, the LSPLR@ARFI-IIT and SOLIROM portals in Iași, the RELATE and TEPROLIN portals at ICIA Bucharest, and the digital format of the Romanian language dictionary, currently undergoing redesign, signify the successful collaboration among researchers and professors in Bucharest, Iași, Cluj, Timișoara, Craiova, and other university centres in the field of Digital Humanities. And because this field is in continuous need of young researchers, Dan Cristea highlighted, the master programme in Computational Linguistics at “Alexandru Ioan Cuza” University in Iași, which produced over 20 generations of graduates since 2001, alongside doctoral studies in the same field. He also mentioned the doctoral school of Computers and Information Technology at the Technical University of Cluj-Napoca and the doctoral school of Mathematics and Computer Science at “Babeș – Bolyai” University in Cluj. These programmes complement main centres in Bucharest, including the doctoral SCOSAAR school led by the “Mihai Drăgănescu” Institute of Research in Artificial Intelligence, the master programmes in Computational Linguistics and Digital Humanities at the University of Bucharest, the doctorates at the Faculty of Mathematics-Informatics, and the doctoral school and Laboratory of Electronics, Telecommunications, and Information Technology at the Politehnica University of Bucharest.

In Nicolae-Victor Zamfir’s opinion, the ongoing exploration in the realm of linguistic technology extends also to the evaluation and processing of human emotions. By delving into the intricacies of emotional nuances, a paradigm shift that could redefine our understanding and interaction with technology is on the verge of appearing, paving the way for a more nuanced and empathetic digital landscape. Thus, the use of computers, various techniques, and sophisticated algorithms plays a pivotal role in comprehending and processing language. At the forefront of this technological evolution, what can be placed nowadays is the phenomenon known as ChatGPT. This innovative tool not only facilitates the creation of works but also synthesises content of creditable quality, an opinion shared by Dan Tufiș, as well, who referred to the remarkable progress that is being witnessed in the field of artificial intelligence (AI), with the emergence of various bots such as Bard, ChatGPT 3 and 4, among others. At the core of these advancements lies the utilisation of massive volumes of data encompassing texts, images, sound recordings, and programming languages. However, with the

benefits of AI also dangers may come, which can be hidden within the training data. Consequently, the collection, filtering, classification, processing, storage, and protection of data have become top priorities on an international scale.

That is why, Dan Tufiş mentioned the projects the European Commission have undertaken to address these challenges, placing a particular emphasis on intellectual property rights. Notable among them is the *European Research Infrastructure Consortium* (ERIC), involving a multitude of partners, including ministries, universities, academies, research institutes, large commercial firms, and SMEs. In his opinion, a distinctive feature of these European initiatives is the promotion and enforcement of interdisciplinarity as an essential premise for advancing knowledge and fostering social development. One such example would be CLARIN, the *Common Language Resources and Technology Infrastructure*, also mentioned by Ruxandra Cosma from the Department of German Language and Literature, University of Bucharest, a digital infrastructure offering data, tools and services to support research based on language resources. CLARIN³ provides easy and sustainable access to a broad range of language data and tools to support research in the humanities and social sciences. CLARIN provides access to multimodal digital language data (text, audio, video) and advanced tools with which to explore, analyse or combine these datasets, being a distributed digital infrastructure, with participating centres all over Europe and further afield, which include universities, research centres, libraries and public archives.

Furthermore, Dan Tufiş referred to the developments achieved in the context of language processing regarding the creation of resources for the Romanian language in recent years. Unfortunately, these accomplishments are often underrepresented in catalogues dedicated to language inventories. To further propel the field, initiatives are underway to establish an open marketplace for language modelling tools and applications. In his opinion, the state of computerisation for the Romanian language is more advanced than suggested by these catalogues, which usually record only metadata. Looking ahead, he recommends the organisation of a workshop, perhaps in 2024, specifically focusing on this topic.

A shared perspective was put forward by Corneliu Burileanu, from the Faculty of Electronics, Telecommunications and Information Technology, National University Politehnica of Bucharest, who, with a background of 40 years in language technology dedicated to speech analysis and synthesis, was of the opinion that, from the perspective of the team in Bucharest, one can no longer speak of the Romanian language as under-resourced, as Romanian already benefits from annotated corpora, at least for spontaneous spoken and written

³ <https://www.clarin.eu/content/clarin-nutshell>

language in continuous discourse. Recognition systems are well-developed with very good results. Similarly, the language synthesis aspect of discourse, according to the view of the research team from Cluj, has also achieved notable progress. For both recognition and synthesis, the Romanian language has entered the international circuit, fulfilling the goals set 15-20 years ago.

Burileanu said that the Romanian language now has enough corpora and databases with metadata for thorough analysis, allowing the focus to shift towards other areas, such as: emotion analysis in speech and the separation between truth and lie in speech. A doctoral thesis from *Transilvania* University of Braşov, which focused on speech technology applications, including emotion analysis and vocal signal encoding-decoding, showing great potential, represents another good example in point.

Romanian BERT – RoBERT was already used in sentiment analysis applications, diacritic restoration, text generation (a new version being under development), corpora, training, evaluation, with special applications that include: false news identification systems – involving the development of a small corpus, summarisation – with student contributions, or evaluation of student performance on Moodle (during the pandemic) based on a comparative corpus analysis before and after Covid, clustering and concept extraction – analysis on social networks, conversational agents, can successfully be provided as other examples.

Ruxandra Cosma, from the University of Bucharest, advanced, as a linguist, the benefits of the existing corpora dedicated to digital studies on the Romanian language, by mentioning the CoRoLa project (*Computational Reference Corpus for Contemporary Romanian Language*), initiated in 2014 as a priority programme of the Romanian Academy. Cosma also mentioned that the adoption of the KorAP platform for managing the CoRoLa corpus was made possible through the DRuKoLA project, funded under the Research Group Linkage Programme of the Alexander von Humboldt Foundation, having as partners the Institute for the German Language in Mannheim, the University of Bucharest, the Institute of Research for Artificial Intelligence of the Romanian Academy “Mihai Drăgănescu” in Bucharest, and the Institute of Computer Science of the Iaşi Branch of the Romanian Academy. The corpus comprises a diversity of texts, covering mainly the period from 1989 to present, thus providing an objective representation of contemporary written and spoken Romanian. The corpus is publicly accessible through two interfaces for textual data search and one for audio data search.

According to the site that hosts it⁴, the primary use domains of the CoRoLa corpus include linguistic studies, language modelling for the automatic processing

⁴ <https://corola.racai.ro/>

of Romanian, translation model development, language learning, intelligent and multi-criteria indexing and retrieval of textual and oral information, semantic classification of large volumes of data (text and audio), knowledge extraction from data (text and audio), automatic document summarization, question-answering systems, speech recognition, and synthesis, etc. The development of the corpus adhered to international standards and best practices, reflecting all functional styles of the language (with English-named categories based on text metadata: Imaginative, Science, Journalistic, Law, Administrative, Memoirs, Blogpost). It covers four general domains (Arts & Culture, Nature, Science, Society), further classified into 70 subdomains. The texts are accompanied by metadata and have undergone a processing chain that combines manual computer-assisted pre-processing and fully automatic processing. Cleaning (removing elements that do not belong to the text - headers, notes, titles, table of contents, etc., from documents), verifying the correct use of diacritics and their automatic insertion, when necessary, as well as creating metadata for each document, were accomplished using internally developed processing tools. Subsequently, the data were segmented at the sentence and word levels, morpho-syntactically annotated, lemmatised, and syntactic groups were identified. Additionally, audio files and their translations were aligned (at the phoneme, syllable, and word levels).

Reference was also made by Ruxandra Cosma to Dariah-Be, whose aim is to offer a sustainable portfolio of services enabling digital scholarship in the arts and humanities in Belgium. To put this into practice, partner institutions are encouraged to establish Digital Humanities Research Centres which, together, form a humanities-specific digital ecosystem, offering services both within their own institutions and to other institutions in Belgium and beyond, as mentioned on the project's site⁵. According to one specific section of the site⁶, Dariah-Be offers a number of collaborative databases, which are developed to meet the needs of researchers in the humanities, as well as a corpus management platform whose goal is to enable collaborative management and discovery of digitised textual collections that allows digital humanities researchers to prepare their corpora (consisting of, for example, digitised newspapers and books) for textual analysis. The platform enables researchers to browse and search the digitised collections compiled, cleaned, enriched and managed by the researchers themselves.

Good practices, according to Gabriela Haja, and the results achieved so far demonstrate that, in the field of Digital Humanities in general, and digital lexicography in particular, collaboration between humanists and computer science experts is essential. The Romanian language and culture need, perhaps more than

⁵ <https://be.dariah.eu/>

⁶ <https://be.dariah.eu/service/collaborative-databases>

ever, cultivation and promotion, especially considering the significant degradation of public discourse quality and the pressure of the English language as the *lingua franca* in scientific communication. Digitisation and digitalisation stand out as significant solutions, with education being paramount at all levels of learning. Therefore, a common strategy and strategic plan for the digitalisation of the Romanian language are necessary.

For specialised languages and the scientific language in general, which, following Haja's argumentation, are directly threatened by the current discouragement of publishing scientific works in Romanian, with disastrous repercussions for education, digitalisation can be a solution. Alongside translating major scientific works into Romanian and creating updated specialised dictionaries, digitalisation can play a crucial role. Lexicographic research, flourishing during the Enlightenment, is once again capturing public interest thanks to digitalisation and advanced possibilities for consultation and usage, ensuring a direct impact of research results in the contemporary world.

Steps towards such an accomplishment have already been taken, according to Gabriela Haja, the campaign of "Alexandru Philippide" Institute of Romanian Philology to promote research results and to educate the target audience, through knowledge access platforms that has been made public and online within programmes of the Romanian Academy, Iași Branch, being a successful example in point. Such initiatives refer to activities such as the programme "Academia Altfel", implemented in relation to the national programme "Școala Altfel", and through the organisation of conferences and lectures in collaboration with partners and colleagues from universities in the country and the Republic of Moldova, starting 2021. Moreover, the establishment of the Commission for the Study of the Origin of the People and the Romanian Language (2019), the Commission for the Strategic Development of Scientific Research (2022) and the Commission of Educational Sciences (2023) also prove the constant interest manifested towards a solid strategic effort for making Romanian coherently digitalised.

Moreover, there is also a need for increased visibility. For this, even if not sufficient on their own, events like upcoming conferences can offer excellent opportunities for visibility. One such academic context is provided by SpeD 2024, mentioned by Corneliu Burileanu, the 13th Conference on Speech Technology and Human-Computer Dialogue to be organised by the University Politehnica of Bucharest / Faculty of Electronics, Telecommunications and Information Technology, The Research Institute of the University Politehnica of Bucharest "CAMPUS", Institute of Computer Science of Romanian Academy – Iași Branch, and The Research Institute for Artificial Intelligence "Mihai Drăganescu" – Romanian Academy, Bucharest, under the aegis of The Romanian Academy.

Initiatives such as the CODHUS research centre (*Centre for Corpus Related Digital Approaches to Humanities*), founded in October 2019, by the Faculty of Letters, History and Theology (FLIT) of the West University of Timișoara, can also increase visibility and can, at the same time, contribute to the joint efforts of a coherent vision in the Digital Humanities field in Romania. CODHUS promotes the complementarity between research and digitalisation in order to obtain immediate results with an applicative character in the trans- and interdisciplinary spectrum, as put forward by Roxana Rogobete, Ana-Maria Bucur, and Mădălina Chitez from the West University of Timișoara, as well as mentioned on the Centre's site⁷. Thus, projects like ROGER and AWICNET, develop a methodology for the analysis of academic writing genres at Romanian universities from a contrastive perspective: genres written in Romanian versus genres written in English; ParlaMint-RO, as part of the European project *ParlaMint: Towards Comparable Parliamentary Corpora*, an initiative of the CLARIN research infrastructure, targets the establishment of a strategy for handling parliamentary textual data; while DACRE, aims to fill the gap in applied corpus linguistics research adapted to the Romanian context, using corpus linguistics methods to investigate on and generate support for discipline-specific academic writing. All of them, described in detail in one of the subsections of the CODHUS' site⁸, fuel the contemporary approaches to the field and further support with solutions to the cohesive nature of the strategy, so much put forward by the dialogue in this meeting.

In conclusion, following a rigorous analysis of what has been done so far (of which only a small part has been mentioned on this meeting of the Consortium), what needs to be done, and the difficulties still to be expected (one of them, very important, being the extremely restrictive copyright legislation), many speakers in the meeting opinionated that, despite the progress made so far, Romania urgently needs a national strategy and plan to guide it according to a coherent and long-term behaviour in the near future (until the end of the decade) towards the digitalisation of the Romanian language.

By its statute, Dan Cristea reminded, the Romanian Academy is responsible for cultivating and promoting the Romanian language, as well as proposing recommendations and national strategies regarding the national policy in the fields of science, literature, and arts. The implementation of such a strategy addressing diverse digital layers of language should be carried out in the academic and research institutions in both our Romanian-speaking countries. Consequently, in developing such a strategy, the Romanian Academy should act as a catalyst for

⁷ <https://codhus.projects.uvt.ro/>

⁸ <https://codhus.projects.uvt.ro/proiecte/>

research and development activities of resources and technologies, boosting this field until achieving perfect compatibility with similar initiatives and creations in Europe and worldwide, stabilising the evolution of the Romanian language, and, at the same time, ensuring its protection in the fierce competition with the highly active languages on the internet.

Sharing the same point of view, Elena Isabelle Tamba stated that Romania is still in need of a medium- and long-term strategy and a national plan to guide the digitalisation of the Romanian language, by the coordination, through concrete actions, of the Romanian language policy with the European language policy, until the gap between the situation in Romania and that of the other EU countries will be zeroed. Seen by Tamba as a cultural strategy, digitalisation, promoted through and extended to modern social media approaches, can generate a brand, while a national portal, supplied with resources and links from credible sources, complimented by joint efforts and a common strategy, can become the interface open to public free access.

Developing the idea, Cristea considers that neglecting this field and delaying the development of a visionary strategy could lead to the perpetuation of the chaotic evolution of research in the field, the already observed irregularities in the unrealistic funding of the Digital Humanities domain compared to others favouring the isolation of research and sometimes the development of partially overlapping research themes. Subsequently, the purpose of his intervention was meant to raise awareness among the Romanian and Moldovan scientific communities, as well as decision-makers in the Romanian Academy (in the leadership of the Philology and Literature Section and the Information Science and Technology Section) and the Academy of Sciences of the Republic of Moldova, as well as in Romanian and Moldovan universities, to develop a strategy for the Romanian language that prioritises the development of language digitalisation projects and directs research activities in this direction. Eventually, the implementation of such a strategy would create the conditions for supporting the Digital Humanities domain in Romania and the Republic of Moldova at the level of presidential and governmental administrations.

References

- Chomsky, N. 1956. Three Models for the Description of Language. *IRE Transactions on Information Theory* 2(3):113-124.
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv Computer Science – Computation and Language, preprint arXiv:1810.04805.
- Forăscu, C., Tufiş, D., and Cristea, D. (eds.). 2006. *Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române*, Editura Universităţii „Alexandru Ioan Cuza” din Iaşi.

- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. 2018. Improving Language Understanding by Generative Pretraining. *Open AI preprint (work in progress)*, URL <https://www.openai.com/research/improving-language-understanding-by-generative-pretraining/>.
- Turing, A. M. 1950. Computing Machinery and Intelligence. *Mind* 59(236):433-460.

KEYNOTE SPEECHES

GABRIELA HAJA

Short bio: Gabriela Haja has a PhD in Linguistics from the “Alexandru Ioan Cuza” University of Iași, and is the director of the “Alexandru Philippide” Institute for Romanian Philology, Romanian Academy, Iași branch, where she has the degree of Scientific Researcher I. Her areas of interest are lexicology, Romanian philology, computerization of Romanian linguistic research, history of Romanian literature, ethnology. She has been actively involved in the creation of several important lexicographic works for Romanian language, among which the Romanian Language Dictionary (DA/DLR, 1913-2010), and is a co-developer of the project CLRE – the Essential Romanian Lexicographic Corpus, including 100 dictionaries from DLR Bibliography aligned by entries.

Contact: <https://acadiasi.org/cercetatori/haja-gabriela/>, gabi.haja@acadiasi.ro

Title: Stages of Digitization in Romanian Academic Lexicography

Abstract: The conference started with a brief analysis of the extent to which society is showing a need for (monolingual) dictionaries, based on a study conducted by European lexicographers in 2018¹ and the digitally active WIKI phenomenon. The main observation is that private initiatives and the collaborative way of producing lexicographic works or corpora is more dynamic than the institutional one, managed by specialists in the field. Following this analysis, the author has outlined a SWOT analysis of the situation of digitization of Romanian language resources and tools within the Romanian Academy, inviting this analysis to be completed in order to develop a strategy and a strategic plan for the digitisation of the Romanian language within the institution responsible for "the cultivation of the Romanian language and literature", "the cultivation and promotion of national and universal science and culture", which is, by statute, the Romanian Academy².

The stages of digitization in Romanian academic lexicography, grouped in four phases, were presented: Phase I. Prospective projects on the computerization of the Dictionary of the Romanian Language (DLR), 2003-2007; Phase II. Developed projects. Special applications. 2007-2010; Phase III. Creation of a Romanian lexicographic corpus. European lexicographic network. 2010-2018;

¹ Iztok Kosem et al., *The Image of the Monolingual Dictionary Across Europe. Results of the European Survey of Dictionary use and Culture*, in *International Journal of Lexicography*, March 2019, 32(1):92-114, <https://doi.org/10.1093/ijl/lec022>

² *Statutul Academiei Române*, in *Monitorul Oficial al României*, Anul 189, Nr. 1152, Partea I, 3 decembrie 2021, page 6.

Phase IV. Romanian Academy. Internal collaboration. Dialogue with European partners (AcadLexi). 2019 — till now. Some of the most important results of this process were presented: the Electronic Lexicographic Corpus (CLRE) (<https://clre.solirom.ro>) which includes, among other works, the Romanian Language Dictionary, also called the Academy's Dictionary or the Treasure Dictionary of the Romanian Language, the creation of the Romanian Academy's Digital Tools and Resources for Romanian Language and Literature Platform (SoLiRom) (<https://solirom.ro>) and the organisation, since 2011, of a scientific event specifically dedicated to the computerisation of Romanian lexicographic research³.

The conference ended with a series of final remarks on the importance of interdisciplinary and transdisciplinary research in the domain of Digital Humanities and of the realization of a Strategy and a Strategic Plan for Digitisation in order to safeguard national languages — the Romanian language, in our case — in the context of globalisation and the threat of English, the present day lingua franca, and, last but not least, to ensure immediate and free access of the general public to the results of academic research.

MARIUS POPESCU

Short bio: Marius Popescu is a lecturer at the University of Bucharest, Department of Computer Science, since 2004. He has got a PhD in 2004 with the thesis “Machine Learning Applied in Natural Language Processing”. His domains of interest are artificial intelligence, machine learning, computational linguistics, information retrieval, authorship identification, computer vision. His achievements in these fields include: a new method for automatic word sense disambiguation that was awarded third prize at the Senseval 3 evaluation exercises for word sense disambiguation organized by ACL-SIGLEX in 2004; co-author of the method ENCOPILOT for plagiarism detection which won the first international competition in automatic plagiarism detection in 2009, followed by ranking 4th at PAN@CLEF 2010 and 2nd at PAN@CLEF 2011; a method for authorship analysis which obtained the best results in the author identification task at PAN@CLEF 2012; co-author of learning systems that ranked on the 3rd place in the Native Language Identification Shared Task of the BEA 8 Workshop of NAACL 2013, and on the 4th place in the Facial Expression Recognition Challenge of the WREPL Workshop of ICML 2013.

Contact: http://old.fmi.unibuc.ro/ro/popescu_marius/, popescunmarius@gmail.com

³ AcadLexi – Lexicografia academică românească. Provocările informatizării / Romanian Academic Lexicography. The Challenges of Computerization, <http://philippide.ro/acadlexi>

Title: AD-NLP: A Benchmark for Anomaly Detection in Natural Language Processing

Abstract: Deep learning models have reignited the interest in Anomaly Detection research in recent years. Methods for Anomaly Detection in text have shown strong empirical results on ad-hoc anomaly setups that are usually made by downsampling some classes of a labeled dataset. This can lead to reproducibility issues and models that are biased toward detecting particular anomalies while failing to recognize them in more sophisticated scenarios. In the present work, the author provided a unified benchmark for detecting various types of anomalies, focusing on problems that can be naturally formulated as Anomaly Detection in text, ranging from syntax to stylistics. The author expressed, in this way, the hope that research on Text Anomaly Detection will be facilitated. The presentation also included evaluation and analysis of two strong shallow baselines, as well as two of the current state-of-the-art neural approaches, providing insights into the knowledge that the neural models are learning when performing the anomaly detection task.

RADU ION

Short bio: Radu Ion is a Scientific Researcher II at the “Mihai Drăgănescu” Research Centre for Artificial Intelligence, Romanian Academy – Bucharest. He has got a PhD in Computer Engineering from the University Politehnica of Bucharest with the thesis: *Word Sense Disambiguation Methods Applied to English and Romanian*. During the years he has been an active member of a number of international projects:

- ACCURAT: with the development of an EM algorithm for document alignment and a parallel sentence mining algorithm;
- SIR-RESDEC: where he contributed to the development of an open-domain QA system answering Romanian natural language questions from JRC-Acquis corpus;
- BalkaNet: checking the inter-lingual semantic alignments between synsets of the Princeton English WordNet and the Romanian WordNet, which was developed in the project. This experiment formed the basis on one of the algorithms described in his PhD thesis;
- MetaNet4U: where he tested/validated a multitude of NLP tools (POS taggers, tokenizers, etc.) and resources (annotated corpora, wordnets, etc.) to comply with the project’s requirements of interoperability and standards compliance. MetaShare (<http://metanet4u.eu/metashare.html>) is the portal where all these resources and tools have been uploaded, the RACAI’s node being also administered by him.

Contact: <https://www.racai.ro/en/about-us/racai-staff/radu/>, ion.radu@outlook.com

Title: The Romanian Language in the Age of Standard and Large Language Models

Abstract: We have entered a decade in which large neural networks, with billions of parameters, can be trained in a matter of weeks, and which can deal with multi-modal information in a way that is unprecedented. These networks perform many tasks related to text comprehension, can produce naturally sounding speech or can extract a lot of information from images or videos. Furthermore, the generative neural networks of different types can produce text, images or speech that can fool almost anyone as to who is the source of these productions: human or machine. In this context, the author took a bird's eye view on how the Romanian language is covered by the text encoding and generation of large neural networks (called Standard Language Models, if they have fewer than a billion parameters and Large Language Models, if they have more than a billion parameters). Then, a discussion was open on whether and how Romanian can have a LLM of its own, trained only on Romanian texts.

LIVIU DINU

Short bio: Liviu P. Dinu is a professor at the University of Bucharest, Computer Science Department, the director of Human Language Technologies Research Center, and member of Computer Science and Interdisciplinary Doctoral Schools. His main research is in Computational Linguistics and Natural Language Processing. He has got a PhD in 2003 with Solomon Marcus as his supervisor, and in 2014 he defended his habilitation thesis entitled "Similarity and Decision Problems in Computational Linguistics". In 2007 he received "Grigore C. Moisil" Prize, awarded by the Romanian Academy (for 2005). He has published 2 books, 8 chapters in books, over 170 papers in journals and conferences proceedings, has initiated and managed 13 national and international R&D projects and was involved in other 14 R&D projects. He has also initiated in 2020 a master program in Natural Language Processing at University of Bucharest.

Contact: <https://nlp.unibuc.ro/people/liviu.html>, ldinu@fmi.unibuc.ro

Title: Computational Approaches in Historical Linguistics

Abstract: Natural languages are living eco-systems, they are constantly in contact, and, as a consequence, they change continuously. Traditionally, the main Historical Linguistics problems (How are languages related? How do languages change across space and time?) have been investigated with comparative linguistics instruments. The main idea of the comparative method is to perform a property-based comparison of multiple sister languages to infer properties of their

common ancestor. It is a time-consuming manual process that required a large amount of intensive work.

The author proposes here computer-assisted methods for identifying cognates, for discriminating between related words, and for protoword reconstruction. The identification of cognates is a fundamental process in historical linguistics, on which any further research is based. Even though there are several cognate databases for Romance languages, they are rather scattered, incomplete, noisy, contain unreliable information, or have uncertain availability. A comprehensive database of Romance cognates and borrowings, based on the etymological information provided by the dictionaries (the largest known database of this kind, to the author's best knowledge) is introduced. Pairs of cognates between any two Romance languages are extracted, by parsing electronic dictionaries of Romanian, Italian, Spanish, Portuguese and French. Based on this resource, a strong benchmark for the automatic detection of cognates is proposed, by applying machine learning and deep learning-based methods on any two pairs of Romance languages. Beside the largest database of this kind, it is also shown that automatic identification of cognates is possible with accuracy averaging around 94% for the more difficult task formulations. Methods for automatic discrimination between related words (cognates vs loanwords, inherited vs loanwords, derivative vs cognates) are then proposed. Finally, a methodology for protoword reconstruction and missing Romanian cognates reconstruction is proposed. Given words in Romance modern languages, the task is to automatically reconstruct the Latin proto-words from which the modern words evolved. The method was applied for producing related words based on sequence labeling, aiming to fill in the gaps in incomplete cognate sets in Romance languages with Latin etymology (producing Romanian cognates that are missing).

RADU TUDOR IONESCU

Short bio: Radu Tudor Ionescu is a professor at the University of Bucharest (Romania). He completed his PhD at the University of Bucharest in 2013. He received the 2014 Award for Outstanding Doctoral Research in the field of Computer Science from the Romanian Ad Astra Association. His research interests include machine learning, computer vision, image processing, medical imaging, computational linguistics and text mining. He published over 120 articles at international peer-reviewed conferences and journals, and a research monograph with Springer. He received the “Caianiello Best Young Paper Award” at ICIAP 2013 for the paper entitled “Kernels for Visual Words Histograms”. Radu also received the “Young Researchers in Science and Engineering” Prize for young Romanian researchers, and the “Danubius Young Scientist Award 2018 for Romania” by the Austrian Federal Ministry of Education, Science and

Research and by the Institute for the Danube Region and Central Europe. Together with other co-authors, he obtained good rankings at several international competitions: 4th place in the Facial Expression Recognition Challenge of WREPL 2013, 3rd place in the NLI Shared Task of BEA-8 2013, 2nd place in the ADI Shared Task of VarDial 2016, 1st place in the ADI Shared Task of VarDial 2017, 1st place in the NLI Shared Task of BEA-12 2017, 1st place in the ADI Shared Task of VarDial 2018, 1st place in the ACM Multimedia 2023 Computational Paralinguistics Challenge (ComParE) on request and complaint detection.

Contact: <https://raduionescu.herokuapp.com/>, raducu.ionescu@gmail.com

Title: Recent Text and Audio Resources for the Romanian Language

Abstract: The talk presented two recent resources for the Romanian language. First, the author introduced a novel Romanian Clickbait Corpus (RoCliCo), comprising 8,313 news samples, which are manually annotated with clickbait and non-clickbait labels. Then, there were presented preliminary results with various machine learning methods. Among the considered models, there has been a novel BERT-based contrastive learning model that learns to encode news titles and contents into a deep metric space. Next, the author referred to RoDia, the first dataset for Romanian dialect identification from speech. RoDia includes a varied compilation of speech samples from five distinct regions of Romania, covering both urban and rural environments, totalling 2 hours of manually annotated speech data. Along with RoDia, a set of competitive models to be used as baselines for future research has been introduced. Moreover, it was presented empirical evidence showing that Automatic Speech Recognition on dialectal speech is more challenging.

CHAPTER 1

DIGITALISATION OF CULTURES

DIGITALISATION OF CULTURES. PREPARATIONS FOR THE ADOPTION OF A STRATEGY FOR THE ROMANIAN LANGUAGE

DAN CRISTEA¹, GABRIELA HAJA²,
ELENA ISABELLE TAMBA², DAN TUFIS³

¹*Institute for Computer Science, Romanian Academy – Iași branch*
dan.cristea@acadiasi.ro

²*“Alexandru Philippide” Institute of Romanian Philology,
Romanian Academy – Iași branch*
gabi.haja@acadiasi.ro, isabelle.tamba@acadiasi.ro

³*“Mihai Drăgănescu” Research Centre for Artificial Intelligence,
Romanian Academy – Bucharest*
tufis@racai.ro

Abstract

This essay, placed in the context of the ConsILR-2023 conference, which also planned a meeting of the *Informatization Consortium for the Romanian Language*, out of which the name of this series of conferences emerged, aims to revive the interest for a union of linguists and computer scientists concerned with the Romanian language in Romania and the Republic of Moldova. After an exposition of the premises that motivate such a concern, we propose a distinction between the terms *digitisation* and *digitalisation* of a nation's culture. The field of *Digital Humanities* (DH) is then briefly presented, with a focus on the area of language processing, insisting on some major achievements of this field of research. A number of significant bodies (associations and consortia) operating at European and global level and a number of research centres known for their significant results in the field are then listed. We then present, without pretending to have exhausted the subject, some of the Romanian and Bessarabian research teams active in the field of digitalisation of the Romanian language, as well as some of their significant achievements. Before formulating the conclusions, the most significant university level education centres training experts capable of contributing to the development of the DH areas in our two countries as well as several series of scientific events in the field are reviewed.

1 Background

All over the world and permanently throughout human history, the transmission of information through symbolic signs has always been the basis of society's progress. For the history of the last seven centuries, the book is the storehouse of knowledge and expression of literary art, the shape and appearance of the object book having had a very important role in the history of mankind. This perspective changed the denotative meaning (cf. DA^d in CLRE¹, s.n. *carte*) of a Romanian saying, since the "book (of education)", "manual", replaced the meaning of "evidential document, (royal) document" of the word *carte* (*letter, epistle*). Thus, *Ai carte, ai parte*², say the Romanians from the middle of the 19th century till our days, urging their children towards school and education.

The written letter, i.e., the text, has known in the course of human history several supports capable of preserving it and transmitting it to others: clay, papyrus, paper, and more recently, tape and magnetic disks, optical devices as well as silicon chips. The middle of the 15th century was the origin of the democratisation of access to information and education, together with the discovery of movable printing and the publication of increasingly large print runs of books. Gutenberg's Bible (a. 1455), edited in this way, opened the way to less expensive printing and represented, for Europe, the birth of the book as a symbol of knowledge. Rotary printing and offset technology revolutionised the printing industry and favoured the generalisation of access to information. Since the fifties of the last century, the digital format began to spread, able to preserve the written letter and, therefore, the accumulation of knowledge of mankind in a much more compact and economical way than paper, although, perhaps, at the same time even more perishable than this. Simultaneously with the appearance of digital media, screens were also largely spread, i.e., devices capable of temporarily reflecting the digitally recorded text or image, so that they can reach the reader's eye, a complication that the paper book does not need, because the storage medium coincides with the display.

Informatics, the science that underlies the developments of the digital society, has found ways to facilitate people's access to the written text, stored in digital format, without requiring any prior, expensive or tedious preparation. For example, in just one year, 2017, the United States Library of Congress recorded more than 110 million website hits, with more than 503 million pages read and

¹ Academia Română, *Dicţionarul limbii române* [DA] (1913–1949), ediţie digitală anastatică (<https://dlr1.soliorom.ro/>) în CLRE. *Corpus lexicografic românesc electronic* (Romanian Academy, *Dictionary of the Romanian language*, digital anastatic edition in CLRE. *Electronic Romanian lexicographic corpus*), 2021–2022.

² Approximately: *Knowledge is power*.

2.7 million pages downloaded, compared to nearly 1.9 million of visitors in situ (2019-2023 Library of Congress Strategic Plan)³. Recent statistics⁴ show that the information stored in digital format today is 64 zettabytes, 90% of it having appeared in the last 2 years. This invasion of society by the digital environment, in addition to the many benefits, has also produced many exaggerations as well as many harmful effects.

Immune to any influence that would come from the fashion of the moment, so labile, influenceable, diverse and controversial in today's society, we propose to discuss in this essay the need to develop a strategy and a plan in view of receiving the attention of the Romanian Academy (and through this – of the Romanian society) to ensure the preservation of the Romanian written culture and the harmonious development of the Romanian language, offering, in the same time, ways of protection against the threat of dissolution in the widely used languages of the planet, so that our language will preserve its identity and specificity, while also retaining its function of integration among the long-standing global technological trends. We are neither original nor trailblazing in writing this essay and we do not aim to cover all the details, our study is only a prolegomenon and an argument in favour of the fact that the moment to act in order to develop such a strategy cannot be delayed any longer, without exposing the Romanian language and culture to major dangers.

Lovers of reading know the overwhelming pleasure of holding a book in their hands, tracing the letters on the printed pages, touching and turning the paper sheets, and even savouring the scent of the printing ink. Each of them has a room or more at home with walls lined with books; books are everywhere around us, on the desk, on the bedside table, in the living room, on the windowsill, in the kitchen, on the floor, we take them with us in our backpack or suitcase when we travel, we lend them to friends (and sometimes we even get them back...). A parallel can be drawn between online education and reading from a screen, as opposed to classroom teaching and reading from paper. The pandemic we recently went through made it necessary to approach, to a much wider extent than before, the distance teaching process, without proving that this means of teaching is superior to the classic one. Direct eye contact between teacher and class, spontaneous dialogue and physical interaction are preferable to online teaching, the same as a visit to a botanical garden is much more than flipping through a botany book. And yet, online teaching can be beneficial under certain conditions, for example, when a university cannot afford to physically bring a renowned professor into the classroom, while online interaction involves minimal costs. Or, even in the

³ https://www.loc.gov/static/portals/strategic-plan/documents/LOC_Strat_Plan_2018.pdf

⁴ <https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/>

classroom, we can use digitised didactic resources and teaching tools, without limiting ourselves to these. The virtual is much more than nothing, but it is less than the real.

In the same way, we can say that reading on classic media is superior to that on electronic media. Pupils or students learn better from books than from screens (Baron 2016; Alexander and Singer, 2017); in terms of the advertising industry, on the other hand, there is a competition between paper and digital advertising⁵, to the detriment of the printed one (Richter, 2019). The digital format has many advantages for preserving information, complementing, but not cancelling, the classic format. We will detail it in the following.

2 What Does the Digitalisation of a Culture Mean?

It is possible that the idea of digitalisation of culture may seem violent, perverse, deviant in the eyes of some people, or be seen as a simple tool in the eyes of others. Neither one nor the others are right. On the one hand, because digitalisation does not invade our lives by turning us into avatars or robots, and on the other hand, because it is much more than a tool.

By *digitalisation* of a culture, implicitly of written texts, we understand not only the placement on electronic media of faithful copies of cultural creations, but also the provision of access facilities to these productions, in new ways, which the original, physical objects cannot assure to the culture consumer. Unlike the very close term *digitisation*, which denotes the change of the support on which a piece is recorded, from physical to digital, digitalisation also includes a change of paradigm, of optics, regarding the manipulation of works. As such, it includes digitisation as only one of the processes with which it operates.

The digital format of a dictionary can allow you to search for information by multiple criteria and get the answer in fractions of a second. Given the digital format of a dictionary, for example, we could issue a request like: "give me the list of most frequently used 100 Romanian feminine nouns with Turkish etymology". Even more, the research in recent years in the field of natural language processing has found solutions for the identification of entities of interest, which could be unknown to a user (at a certain moment). Such searches are based on contexts of use or discursive combinations. The way of access to cultural creations is extremely important, because, often, novice art consumers are not aware of the existence of the object they are looking for, therefore having to be guided by analogies and preferences, gradually narrowing their search area, so that, in the end, they can choose, from a number of similar exposures, those (or

⁵ Print vs digital advertising, in "Newsworks", 7 July 2020.

only the one) that satisfy (satisfies) their personal requirements. It is not unusual to visit a gallery with the main purpose of contemplating in person a very beloved painting, which we knew only from reproductions, but, in many cases, we go there to discover other art objects that we did not know about or that show similarities with a work or a painter or a current that we love.

The digital format of art copies the original, thereby being inferior to the original. But, by copying and distributing it (with the author's consent, of course), it raises the possibility of being known and or "tangible" to anyone. In the rest of this section, we will take the example of a digital dictionary to display some of the advantages of the digital format over the classical one:

- **Dissemination:** the digital format infinitely amplifies the access possibilities of interested readers, by offering extremely easy and universal access. The readers are not more constrained to possess the dictionary in their own library or to go to a public one for searching it: anyone can access it from the laptop or even the portable telephone.

- **Manipulation:** the digital format removes the need for physical handling of printed dictionaries. Usually, good dictionaries are huge and, therefore, extremely voluminous⁶. A simple digital device is sufficient to browse through a very large dictionary.

- **Multiplication:** the digital support corresponds to a multiplication of the object, because in its digital form it can be accessed simultaneously by an unlimited number of users.

- **Amplification of the content:** a dictionary entry can be linked to multiple other objects, transgressing or not the original genre; for example, a citation exemplifying a sense of a word can be linked in the original book from where the citation was extracted, thus revealing a much larger context; a click on the title word can produce its pronunciation or access to an image, to web pages, to augmented reality, to music, etc.

- **Diversity of access paths to the content:** the sequential browsing of the dictionary entries when searching a word, available in the case of the classic format (sometimes helped by fine relief markings cut by the book publisher on the opposite side of the spine of a very voluminous dictionary, displaying the letters of the alphabet, thus facilitating a partial direct access) is, usually, replaced by a direct lookup, when the searched-for word is known. But, sometimes, the users are unable to remember the terms they are looking for⁷. In these cases, an approximate search can be initiated by using either a combination of criteria or a

⁶ The original edition of the Dictionary of the Academy counts 36 volumes, while the 2010 anastatic edition has 19 large volumes.

⁷ As in the case of the tip-of-the-tong problem.

successive narrowing of a search front until the desired term is brought to the centre of attention.

- **Preservation of the digitised object:** the physical format is subject to erosion and deterioration over time. Even though the digital media itself is volatile (because any electronic information storage material has a limited life), backups, periodically refreshed, ensure a virtually unlimited life of the digital object.

- **Dictionary building:** the digital format of a dictionary also obliges the lexicographers to adopt a completely diverse technology for building it. The classic and time-consuming hunt of citations in the recommended bibliography, their manual indexing, their inventory on paper files and their selection to include the most representative for exemplifying senses are set aside. Guided searches of the terms on huge collections of written texts can be placed at the disposal of lexicographers and their work can be facilitated by intelligent frontends to select representative examples. Moreover, word-spies technologies can be implemented to permanently survey growing collections of written texts to select new dictionary entries or new senses of words, or to signal deterioration of uses of a word, thus proposing to lexicographers modifications to be taken into consideration for future editions (Cristea, 2010; Cristea, 2011).

3 The Field of Digital Humanities

The term used in English to name the field of research that we are dealing with in this essay is *digital humanities*⁸ (DH). The field, at the intersection of the humanities and computer science (or information technology), is in a continuous and expansive development, including computer applications dedicated to the preservation and exploitation of cultural heritage and providing humanists with sophisticated research tools capable of opening new avenues of exploration. Of major importance in this direction is the recognition that the printed document is no longer the primary medium for storing knowledge and disseminating cultural production. But subject to the study of this field are not only textual objects, because here you can find 3D architectural reliefs and digital reconstructions, musical scores, scanned pages from heritage books, virtual museums, etc.

Digital humanities cover a variety of applied research topics, from building and managing online collections of primary resources (mostly textual) to discovering regularities in large cultural data sets (cultural data mining).

In the area of text processing, we have on the one hand, minimally processed data: collections of texts (electronic corpora, general or grouped by themes, dictionaries and lexicons, images of original pages, etc.) that reproduce in

⁸ Which has a sweet Romanian translation: *umanioare digitale*. A less concise variant also circulates – *ştiinţe umaniste digitale*.

electronic format, without any intervention, the physical originals, i.e., the results of digitisation operations. But often, the original data are augmented with annotations performed by experts (or automatically, when the technology has reached a level of accuracy comparable to that of human experts) and then they become: lexically, syntactically, semantically annotated corpora, grammars and treebanks (i.e., large collections of syntactic trees), lexical-semantic structures of the vocabulary of a language (such as wordnets), dictionaries indexed in themselves or in other resources (corpora, wordnets, etc.) or numerical data obtained from textual data (frequency tables, models of language, statistical, vectorial or neural representations).

On the other hand, there are the computer applications that process these data: systems for morpho-syntactic annotation, automatic translation, summarization, dialogue agents (like chatbots), semantic disambiguation of words in context, anaphora resolution, deciphering the emotions expressed in the text, syntactic parsing (generation of syntactic trees of phrases), etc. Other types of applications are interactive in nature, helping the human user to perform certain operations on texts or other types of digital objects, faster and better: assisted translation, generation of comparative critical editions, assistance systems in digital drafting and publishing, cultural analysis, etc.

Modern people are in a hurry. Always short of time, they look for electronic means of information, which are conveniently at their fingertips. Young people visit libraries less and less, they have become more superficial than their grandparents of half a century ago or even than their parents of 25 years ago, although they are probably informed in more fields than them. The modern man must be able to find and process very diverse information in a very short time, without carrying with them a vast library. Scientists mainly get their information from the Internet, a laptop being the only tool they need in order to write a very well documented article. Only when the research is tributary to original artefacts, which must be scrutinised to decipher hidden details, and which might require investigations to the field, to an archaeological site or to a museum, the portable computer is no longer enough. Even so, having alongside a transposal of the exhibits into digital format can provide unsuspected sources of inspiration, by discovering similarities, while cataloguing them in hierarchies or comparing them against objects sometimes located on the other side of the globe.

In the field of linguistics, computer-aided research opens huge perspectives, due to the speed of access to sources, the possibilities of searching according to multiple criteria, selective capacities, classification and indexing. In the work of a lexicographer, the operation of identifying citations to exemplify a meaning of a dictionary entry can today be greatly accelerated by occurrence-finding programs, capable of extracting text sequences in which the word of interest

appears, surrounded by lexical elements that ensure the semantic specificity desired by the lexicographer. No institute of lexicography in the world uses the classic filing system anymore, with the manual sorting of the files transcribed by hand from the bibliographies. Institutes of lexicography, lexicology, and language studies have, for the most part, changed their profile. Which is not to say that they have given up on theoretical linguists; these experts continue to be the backbone of research. But they feel better now when having them, shoulder by shoulder, specialists in computational linguistics, natural language processing, or digital humanities. In all these places, whether they are in Princeton (the birthplace of Wordnet⁹), in Nancy (in the *Laboratoire d'Analyse et Traitement Informatique de la Langue Française*¹⁰) or in Mannheim (in the *Leibniz-Institut für Deutsche Sprache*¹¹), computer technology is present and helps in the operations of editing dictionaries, building corpora, annotating primary documents, linguists and computer scientists working together in mixed teams. At the same time, the results of their work are widely distributed on the Internet for the intellectual profit of all those interested, always respecting the copyright of the owners of the texts.

Theoretical linguistics – and experimental linguistics even more so – calls on the computer to select and organise linguistic material, even to abstract and formalise a grammar or to verify the linguistic coverage of a proposed form for a grammatical rule. Computational support is more and more required, including in grammatical analysis or language generation, although today these are primarily achieved by means that are no longer anchored in symbolic constructions, but in statistical or diffuse representations, such as neural models.

Now new syntagms characterise directions of study in which more and more significant research capacities are concentrated. Among them – *Open Linguistic Linked Data*, announcing the concern for creating links between linguistic resources of different types, thereby aiming to create synergies that complement and amplify the evocative force of resources considered independently, with openness to knowledge from auxiliary fields, at the same time campaigning for unrestricted access to knowledge and information.

4 European and Global Organisations

The beginnings of this field must be correlated with the appearance of the computer itself, i.e., the 1950s, when the computer, even as primitive as it was at that time, was used by researchers to search and count words in texts or for indexing documents. The first journal in the field should probably be considered

⁹ <https://wordnet.princeton.edu/>

¹⁰ <https://www.atilf.fr>

¹¹ <https://www.ids-mannheim.de/>

Computers and the Humanities (1966). In the following decades, specialised organisations appear in certain subfields, such as *The Association for Literary and Linguistic Computing* (ALLC) or *The Association for Computers and the Humanities* (ACH), *European Language Resources Association* (ELRA), *Special Interest Group on Under-resourced languages* (SIGUL), the *UNESCO Intergovernmental Information for All* (IFAP) program and the *Interregional Library Cooperation Centre*.

An important step in the development of the field dedicated to text processing was *Text Encoding Initiative* (TEI – 1987), which defined a standard, resumed in 1994 and kept up to date since then¹², making compatible digital textual productions between different languages and lexicographic traditions. Librarians and archivists have an important say in DH projects, with their work covering digital acquisition and stewardship, essential to the preservation, promotion, and access to digital collections. Thus, among the significant projects of the DH field, it is worth mentioning those of the big libraries in the world (at federal, national or university level), among which – the projects of the Romanian Academy Library or the university libraries in the country, which are in the process of creating digital libraries.

Seemingly paradoxical, the preservation of linguistic identity is a priority in today's globalised society. Despite the tendency that some hasty commentators might insinuate, who might confuse globalisation with uniformity, a lot of international initiatives aim to integrate diversity, encourage national specificity and safeguard endangered languages. All over Europe and the world, legislative strategies and policies are being initiated that encourage theoretical and applied research projects focused on the digitisation of languages. In Europe, an important role is played in this direction by the *European Federation of National Institutions for Language* - EFNIL¹³ (the Romanian Academy being a member of this federation since 2019), which brings together over 40 members from 30 countries, representatives of the major academic linguistic organisations and other national language bodies of the Member States of the European Union and associated countries, with the role of overseeing the monitoring of the official language (or languages) of their countries and providing advice on the use of languages or the development of language policies. EFNIL is also an advisory body for the European Commission, regarding the European language policy, a partner in the *European Language Equality* project of EU¹⁴, which aims to develop a unified European strategy to achieve equality for the promotion of languages in Europe

¹² <http://www.tei-c.org/index.xml>

¹³ <http://www.efnil.org/>

¹⁴ <https://european-language-equality.eu/>

by 2030¹⁵. Another structure, representative of its initiatives and efforts regarding the digitisation of lexicographic research, established in Europe, but with members from all over the world, is the *European Association for Lexicography* – EURALEX¹⁶, the most prestigious professional association in the field of lexicography, in the broad field of language sciences and in related fields.

At the European level, this genre of concerns has a long tradition. Since the 1960s, institutes and research laboratories dedicated to this direction have been created. Exemplary is the activity of the ATILF laboratory¹⁷, subordinate to the CNRS and the Université de Lorraine, which has enjoyed national and European investments, with outstanding results, among which we note: *Le Trésor de la langue française informatisé* – TLFi, the *Frantext* French language corpus, the collection of resources and French language processing tools *Outils et Ressources pour un Traitement Optimisé de la LANGue* – ORTOLANG, *Dictionnaire du Moyen Français* – DMF, *Dictionnaire Étymologique Roman* – DÉRom (two volumes published, ongoing project, in which specialists from Romania also collaborate), tools of advanced consultation of the *Dictionnaires de l'Académie Française* (editions 4, 8 and 9), *Français, Langue Étrangère Universitaire: Ressources et Outils Numériques* – FLEURON (dedicated to students who wish to initiate, from distance, the study of the French language), and still others.

The European Parliament resolution of 11 September 2018 on language equality in the digital age urges Member States to propose effective ways of supporting their own language, to develop comprehensive policies in language-related areas, to allocate resources and use appropriate tools to promote and facilitate linguistic diversity and multilingualism in the digital domain. The resolution emphasises the joint responsibility of the EU and Member States, as well as universities and other national public institutions, in taking measures to preserve their languages in the digital world and in developing databases and translation technologies in all EU languages, including in languages with fewer speakers.

In Europe and the world, at the academic level, there is an obvious concern about the digitization of resources and electronic access to textual information¹⁸ (Tamba, 2022). In 1982, André Danzin together with Nobel Laureate Ilya Prigogine (1982) described the role of science as a trigger for societal change and noted that planning research and setting precise goals can lead to the diminution of innovation and creativity. Ten years later, the same André Danzin, rapporteur

¹⁵ Thus, the 2023 EFNIL Conference had as its theme: By the People, for the People: Online Dictionaries, Language Portals, and the Role of Language Users.

¹⁶ <https://euralex.org/>

¹⁷ <https://www.atilf.fr>

¹⁸ <https://www.atilf.fr/laboratoire/presentation-francais>

of the European Commission, formulates an objective for the field of European language policies, to counter the potential threat of loss of linguistic identity in the digital age, namely to create electronic resources and tools for each language and culture in particular: "In the electronic age, it is essential to the survival of a language that it be used in electronic information systems." (Danzin, 1992). But his conclusions, formulated in 2000, are still more general, he noting that the process of globalisation in the field of economics is an irreversible one, and the way by which a society, like the French one, which he analyses in his study, could pass the "globalisation shock" was only through an evolution in accordance with the self-image, i.e., in harmony with one's own identity¹⁹ (Danzin, 2000).

5 Digitalisation of the Romanian Language

In Romania and the Republic of Moldova, there have been concerns in the direction of DH for several decades already, through the efforts of some research and education teams from Bucharest, Iași, Cluj, Chișinău, but also from Craiova, Timișoara and other university centres, and sometimes, even outside the Romanian language area, through researchers from the diaspora.

More important than the achievements themselves, only few of them being mentioned below, it is worth noting that Romania and the Republic of Moldova have specialists capable of carrying out activities in this field at a high level of expertise, thanks to the existence of university schools of computational linguistics and natural language processing, at the universities of Iași²⁰, Bucharest²¹, Cluj-Napoca²² and Chișinău²³. Although all these high education schools at the master and doctoral levels are attached to faculties and/or universities with technical profiles (computer science, computer engineering), it is important to note that many of their graduates have a humanistic background

¹⁹ <https://gallica.bnf.fr/ark:/12148/bpt6k442181r.image.r=J%C3%A9sus.f25.pagination.langFR>

²⁰ The Master's Degree in Computational Linguistics from the "Alexandru Ioan Cuza" University in Iași has produced, since its establishment in 2001, 20 generations of graduates, and in the Doctoral School of Informatics, 13 PhD thesis with topics in the field have been completed.

²¹ The SCOSAAR doctoral school led by the Artificial Intelligence Research Institute "Mihai Drăgănescu" of Romanian Academy, the Natural Language Processing master's degree and the doctoral series from the Faculty of Mathematics and Computer Science of the University of Bucharest, the Doctoral School from the Faculty of Automation and Computers, as well as the Doctoral School and the Laboratory of Electronics, Telecommunications and Information Technology (SD-ETTI) from the Faculty of Electronics and Telecommunications, the last two from the National University of Science and Technology POLITEHNICA Bucharest.

²² The Doctoral School of Computers and Information Technology of the Faculty of Automation and Computers of the Technical University, as well as the Doctoral School of Mathematics and Informatics of the Faculty of Mathematics and Informatics of Babeș - Bolyai University.

²³ At the Technical University and the State University of Moldova.

(for example, the Master's degree from Iași is also open to students with a degree in Bachelor of Arts). The bond between computer scientists and linguists was forged through an effort lasting more than two decades and in which joint projects and scientific events organised and carried out jointly by specialists from both fields were relevant.

Among the projects jointly developed by computer scientists and linguists, we mention only a few of those carried out in Iași, Bucharest and Cluj-Napoca: eDTLR: *the Thesaurus dictionary of the Romanian language in electronic format* (2007-2010) – (Cristea *et al.*, 2007; Cristea *et al.*, 2011; Haja, 2017a), DRuKoLa: *Comparing German and Romanian languages through corpus linguistics technologies* (2013-2017) – (Cosma *et al.*, 2016), CoRoLa: *the Corpus of the Contemporary Romanian Language* (2013-2017) – (Cristea *et al.*, 2019; Tufiș *et al.*, 2019; Gîfu *et al.*, 2019; Barbu Mititelu *et al.*, 2018), ReTeRom: *Resources and technologies for the development of human-machine interfaces in the Romanian language* (2017-2020) – (Bibiri *et al.*, 2020), DeLoRo – *Deep Learning for Old Romanian Language* (2019-2022) – (Cristea *et al.*, 2021; Cristea *et al.*, 2023), SemRO – *Computational Representation in Linguistic Big Data Processing* (2021-2022), E-READ: *Evolution of Reading in the Age of Digitization* (2014-2018) – (Bădulescu and Cristea, 2020), MappingBooks – *Let's jump in the book!* (2014-2016) – (Cristea and Pistol, 2014), CLARIN - *Common Language Resources and Technology Infrastructure* (2008-2011) – (Cristea and Pistol, 2009), ENeL - *European Network of e-Lexicography* (2013-2017), SSPR - *Syntactic Analyzer (for Romanian) based on semantic knowledge*, ACCURATE-RO – *Analysis and evaluation of Comparable Corpora for Under Resourced Areas of machine Translation* (Skadina *et al.*, 2010), STAR – *Automatic Translation System for the Romanian Language*, CLARIN-RO – *Infrastructure for interoperable linguistic resources for the Romanian language*, SIR-RESDEC – *Question-Answering System in Romanian and English with Open Search Spaces*, METANET4U – *Enhancing the European Linguistic Infrastructure*, etc.

Regarding the institutional projects carried out within the institutes with a humanistic profile, it is worth mentioning: Dicționarul limbii române în format electronic. Studii privind achiziționarea (The Romanian language dictionary in electronic format. Acquisition Studies) (2003-2005) – (Haja, 2017b), Resurse lingvistice în format electronic. Monumenta linguae Dacoromanorum. Biblia 1688. Pars VII. Regum I, Regum II – ediție critică și corpus adnotat (Language Resources in Electronic Format. Monumenta linguae Dacoromanorum. Bible 1688. Pars VII. Regum I, Regum II – critical edition and annotated corpus)²⁴ (2006-2007) – (Haja, 2017c), CLRE – *Corpus lexicografic românesc esențial*

²⁴ <https://biblia1688.solirom.ro>

(Essential Romanian Lexicographic Corpus) (2010-2013) – (Tamba, 2017), DLR – Dicționarul limbii române. Ediție a doua revizuită și adăugită (Dictionary of the Romanian Language. Second Revised and Added Edition) *et al.*

Among the events with joint organisation, we mention: the 16 schools of the EUROLAN series (held between 1993-2023)²⁵, the ConsILR conference series (held between 2001 and 2023, the current one being the 18th edition²⁶), some of the 12 editions of the colloquium series *Lexicografia academică românească. Provocările informatizării*²⁷ (*Romanian Academic Lexicography. The challenges of informatisation*), the series of meetings of the members of the joint German-Romanian project DRuKoLa, the series of workshops organised as satellite events of the ConsILR conferences (in general, with inserts in the projects), the Iași – Chișinău collaborations in which are usually involved linguists and computer scientists from both countries, the workshops to call for collaboration and to popularise the eDTLR project²⁸, the "Solomon Marcus" symposium series organised by the Faculty of Mathematics and Informatics in Bucharest²⁹, the SpeD conference series³⁰ on speech processing and human-computer dialogue, organised by the National University of Science and Technology POLITEHNICA Bucharest, as well as many others.

In line with the current European language policy, an extensive digitisation process of philological research is underway in our country, aimed at facilitating access to the resources created for the Romanian language. Among the results of the permanent collaboration of linguists with computer scientists over the last 20 years in Romania, the following can be mentioned: CoRoLa – the *Corpus of the Contemporary Romanian Language*³¹, CLRE – the *electronic Romanian lexicographic corpus*³², ROCC – the *Romanian Old Cyrillic Corpus*³³, RODICA - *ROmanian DIachronic Corpus with Annotations*³⁴ the treebanks collection for the Romanian language, the LSPLR@ARFI-IIT³⁵ and SOLIROM³⁶ portals from the Iași Branch of the Romanian Academy, as well as the digital format of the

²⁵ See the last and the proceeding editions at: <https://eurolan.info.uaic.ro/2023>

²⁶ <https://profs.info.uaic.ro/~consilr/2023/>

²⁷ https://solirom.ro/publications/COLEXI_2020-2021.pdf, https://solirom.ro/publications/COLEXI_2022.pdf

²⁸ At the faculties of Letters in Iași, Bacău, Suceava, Baia Mare, between 2007 and 2010.

²⁹ <http://nlp.unibuc.ro/events.html>

³⁰ <https://sped.pub.ro>

³¹ <https://corola.racai.ro>

³² <https://clre.solirom.ro>

³³ The Online Old Cyrillic Image Annotation Tool (OOCIAT) can be accessed at: <http://deloro.iit.academiaromana-is.ro/image-annotator/index.php>

³⁴ More details in Gîfu, 2016 and <http://lsplr.iit.academiaromana-is.ro/resources/detail/7>

³⁵ <http://lsplr.iit.academiaromana-is.ro/>

³⁶ <https://solirom.ro/>

Romanian language dictionary, the thesaurus dictionary of the Romanian Academy – currently in a redesign process. Similar to the practices in other European Union states, the resources and tools created in the Romanian Academy, through its institutes with a philological and IT profile, are made available to researchers interested in studying the Romanian language, some of them being open for access to the general public as well.

6 Conclusions

With all the progress made so far, Romania is in acute need of a strategy and a national plan that will guide it in a coherent behaviour in the near future (until the end of the decade) and lasting, in the direction of the digitalisation of the Romanian language. By statute, the Romanian Academy is responsible for the cultivation and promotion of the Romanian language, as well as for proposing national recommendations and strategies regarding the national policy in the field of science, letters and arts. The implementation of such a strategy should be done in the university and academic research institutes of our both countries of Romanian language, after a rigorous analysis of what has been done so far (of which only a small part has been mentioned in this essay), but also of what follows to be done, and the difficulties expected (one of them, very important, being the extremely restrictive legislation regarding the copyright).

We believe that the elaboration of such a strategy of the Romanian Academy regarding the digitisation of our language would act as a catalyst for the research activity and development of resources and technologies, boosting this field in the Romanian language areas, until making it compatible with similar initiatives and creations from Europe and the world, stabilising the evolution of the Romanian language and, at the same time, ensuring its protection in the fierce confrontation with the extremely active languages on the Internet.

On the other hand, neglecting this field and delaying the development of a visionary strategy means perpetuating the chaotic evolution of research in the field, the irregularities already noticed in the realistic funding of the DH field as compared to others, it favours the isolation of researchers and even the development of themes of partially overlapping research.

The purpose of this essay is therefore to sensitise the Romanian and Bessarabian scientific communities, as well as the decision-makers in the Romanian Academy (from the leadership of the Section of Philology and Literature and the Section of Science and Information Technology) and in the Romanian and Moldavian universities, in order to develop a Strategy for the Romanian Language that would prioritise the development of language digitalisation projects and would orient the research activity in this direction.

Last but not least, the implementation of such a strategy would create the premises for supporting the DH field in Romania and the Republic of Moldova at the level of the Presidential and Governmental Administrations.

Placed in the context of the ConsILR-2023 conference, which also planned a meeting of the *Informatization Consortium for the Romanian Language*, out of which the name of this series of conferences emerged³⁷, our desire is to revive the interest for a union of forces of researchers working in linguistics and computer science concerned with the Romanian language.

References

- Alexander, P. A. and Singer Trakhman, L. S. 2017. The Enduring Power of Print for Learning in a Digital World. *The Conversation*, Europe Edition, 4 October³⁸.
- Baron, N. S. 2016. Do Students Lose Depth in Digital Reading? *The Conversation*, Europe Edition, 20 July.
- Bădulescu, D. and Cristea, D. 2020. Reading with the Body and the Bodies of Books. *Philologica Jassyensia*, an XVI, 1(31):13–22.
- Bibiri, A., Boghiu, Ș., Cristea, D., Gădioi, L., Gifu, D., Onofrei, M., Pericică, F. C., Pistol, I. Scutelnicu, A., and Trandabăț, D. 2020. ReTeRom - proiectul 1: CoBiLiRo – Corpus Bimodal de Limbă Română (ReTeRom - project 1: CoBiLiRo – Bimodal Corpus of the Romanian Language). The 3rd ReTeRom Workshop, ConsILR.
- Cosma, R., Cristea, D., Kupietz, M., Tufiş, D., and Witt, A. 2016. DRuKoLa – Towards Contrastive German-Romanian Research based on Comparable Corpora. *Proceedings of the Workshop on the Challenges in the Management of Large Corpora (CMC-4)*, 28 May, Portoroz.
- Cristea, D. 2010. Very large language resources? At our finger! *Proceedings of the Workshop Language Resources: From Storyboard to Sustainability and LR Lifecycle Management*, LREC 2010, Malta.
- Cristea, D. 2011. Romanian Linguistic Resources on Very Large Scale. *Computer Science Journal of Moldova*, vol.19, 2(56):130-145.
- Cristea, D., Cleju, N., Rebeja, P., Haja, G., Coman, E., Vasilescu, A., Marinescu, C., and Dascălu, A. 2023. Bringing the Old Writings Closer to Us: Deep Learning and Symbolic Methods in Deciphering Old Cyrillic Romanian Documents. *Memoirs of the Scientific Sections of the Romanian Academy*, Tome XLVI, Digital Humanities, pages 87-125.
- Cristea, D., Diewald, N., Haja, G., Măranduc, C., Barbu Mititelu, V., and Onofrei, M. 2018. How to Find a Shining Needle in the Haystack. Querying CoRoLa: Solutions and Perspectives. *Revue Roumaine de Linguistique*, 64(3), Publishing House of the Romanian Academy.

³⁷ Created in 2001, simultaneously with the Informatization Commission for the Romanian Language, a commission of the Romanian Academy.

³⁸ Online at: <https://theconversation.com/the-enduring-power-of-print-for-learning-in-a-digital-world-84352>

- Cristea, D., Haja, G., Moruz, A., Răschip, M., and Patrașcu, M. I. 2011. Statistici parțiale la încheierea proiectului eDTLR – Dicționarul Tezaur al Limbii Române în format electronic (Partial statistics at the end of the eDTLR project - Thesaurus Dictionary of the Romanian Language in electronic format). In Rodica Zafiu, Camelia Ușurelu, Helga Bogdan Oprea (eds.) *Limba română. Ipoteze ale variației lingvistice. Actele celui de-al 10-lea Colocviu al Catedrei de limba română (Romanian language. Hypotheses of linguistic variation. Proceedings of the 10th Colloquium of the Department of Romanian Language)*, Bucharest, 3-4 December, vol. I, *Gramatică și fonologie, lexic, semantică, terminologii, istoria limbii române, dialectologie și filologie (Grammar and phonology, lexicon, semantics, terminologies, history of the Romanian language, dialectology and philology)*, Bucharest, The University of Bucharest Printing House, pages 213-224³⁹.
- Cristea, D., Pădurariu, C., Rebeja, P., Scutelnicu, A., and Onofrei, M. 2021. Data Structure and acquisition in DeLoRo. A Technology for Deciphering Old-Cyrillic-Romanian document. In *Proceedings of the 16th International Conference "Linguistic Resources and Tools for Processing the Romanian Language"*, online, 13-14 December, pages 59-74⁴⁰.
- Cristea, D. and Pistol, I. C. 2009. Limba română în perspectiva CLARIN (The Romanian language in the CLARIN perspective). Diana Trandabăț *et al.* (eds.) *Lucrările atelierului „Resurse lingvistice și instrumente pentru prelucrarea limbii române, Iași, 19-21 noiembrie 2008” (Proceedings of the workshop "Linguistic Resources and Tools for Processing the Romanian Language, Iasi - November 19-21")*, Alexandru Ioan Cuza University of Iasi (UAIC) Publishing House.
- Cristea, D. and Pistol, I. C. 2014. MappingBooks: Linguistic Support for Geographical Navigation Systems. Mihaela Colhon *et al.* (eds.) (2014). *Proceedings of the 10th International Conference "Linguistic Resources and Tools for Processing the Romanian Language, Craiova, 18-19 September 2014"*, UAIC Publishing House, pages 189-198.
- Cristea, D., Răschip, M., Forăscu, C., Haja, G., Florescu, C., Dănilă, E., and Aldea, B. 2007. The Digital Form of Thesaurus Dictionary of the Romanian Language. Corneliu Burileanu, Horia-Nicolai Teodorescu (eds.), *Proceedings of SPeD-2007*⁴¹, Ed. Academiei Române, București, pages 195-206.
- Danzin, A. and Prigogine, I. 1982. Are We Getting the Science We Need? A Hard Look at Science as a Trigger of Change. *UNESCO Courier*⁴², February 1982, pages 4-9.
- Danzin, A. 1992. Towards a European Language Infrastructure, CEC Doc. 54210/92.
- Danzin, A. 2000. Economie: La culture française est-elle inadaptée? *Études* 2000/7-8(393):23-36.

³⁹ Online at: <http://metanet4u.eu/publications/partial.pdf>

⁴⁰ Online at: https://profs.info.uaic.ro/~dcristea/papers/Data%20Structure%20and%20Acquisition%20in%20DeLORo%20-%20a%20Technology%20for%20Deciphering%20Old%20Cyrilic-Romanian%20Documents_final.pdf

⁴¹ Online at: <https://profs.info.uaic.ro/~dcristea/papers/Cristea%20et%20al-SPeD07.pdf>

⁴² Online at: <https://en.unesco.org/courier/february-1982>

- Gifu, D., Moruz, A., Bolea, C., Bibiri, A., and Mitrofan, M. 2019. The Methodology of Building CoRoLa. *Revue Roumaine de Linguistique / On Design, Creation and Use of the Reference Corpus of Contemporary Romanian and Its Analysis Tools*. CoRoLa, KorAP, DRuKoLa and EuReCo / Conception, création et utilisation du Corpus de Référence du Roumain Contemporain et de ses outils d'analyse. CoRoLa, KorAP, DRuKoLa et EuReCo, Ruxandra Cosma and Marc Kupietz (eds.), 64(3):241-253.
- Gifu, D. 2016. Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language. PhD thesis, "Alexandru Ioan Cuza" University of Iași, Romania.
- Haja, G. 2017a. eDTLR – Dicționarul tezaur al limbii române în format electronic (eDTLR – Thesaurus dictionary of the Romanian language in electronic format) (2007-2010). Haja Gabriela (ed). *Lexicografia academică românească. Studii. Proiecte (Romanian academic lexicography. Studies. Projects)*, Iași, "Alexandru Ioan Cuza" Printing House, pages 203-219.
- Haja, G. 2017b. Dicționarul limbii române în format electronic. Studii privind achiziționarea (2003-2005). Haja Gabriela (ed). *Lexicografia academică românească. Studii. Proiecte (Romanian Academic Lexicography. Studies. Projects)*, Alexandru Ioan Cuza University of Iasi Publishing house, pages 181-184.
- Haja, G. 2017c. Resurse lingvistice în format electronic. Monumenta linguae Dacoromanorum. Biblia 1688. Pars VII. Regum I, Regum II – ediție critică și corpus adnotat (Language Resources in Electronic Format. Monumenta linguae Dacoromanorum. Bible 1688. Pars VII. Regum I, Regum II – critical edition and annotated corpus). Haja Gabriela (ed). *Lexicografia academică românească. Studii. Proiecte (Romanian academic lexicography. Studies. Projects)*, Alexandru Ioan Cuza University of Iasi Publishing house, pages 185-193.
- Barbu Mititelu, V., Tufiș, D., and Irimia, E. 2018. The Reference Corpus of the Contemporary Romanian Language (CoRoLa). *Proceedings of LREC 2018*, Japan, pages 1178-1185.
- Richter, F. 2019. The Dramatic Decline of Print Advertising. In *Statista*⁴³, 9 December.
- Skadiņa, I., Vasiljevs, A., Skadiņš, R., Gaizauskas, R., Tufiș, D., and Gornostay, T. 2010. Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, LREC 2010, Malta, 22 May, pages 6-14.
- Tamba, E. I. 2017. CLRE. Corpus lexicografic românesc esențial. 100 de dicționare din Bibliografia DLR aliniată la nivel de intrare și, parțial, la nivel de sens (CLRE. Essential Romanian lexicographic corpus. 100 dictionaries from the DLR Bibliography aligned to entry level and partially to meaning level). Haja Gabriela (ed). *Lexicografia academică românească. Studii. Proiecte (Romanian Academic Lexicography. Studies. Projects)*, Iași, UAIC Printing House, pages 221-234.
- Tamba, E. I. 2022. The Role of the Institutes of the Romanian Academy in the Digitalization Process of the Linguistic Research. In Željko Jozić, Sabine Kirchmeier (eds.). *The Role of National Language Institutions in the Digital Age. Contributions to*

⁴³ Online at: <https://www.statista.com/chart/20244/estimated-print-advertising-revenue-in-the-united-states>

the EFNIL Conference 2021, Cavtat. Budapest, Nyelvtudományi Kutatóközpont, pages 91-100.

Tufiş, D., Barbu Mititelu, V., Irimia, E., Păiş, V., Ion, R., Diewald, N., Mitrofan, M., and Onofrei, M. 2019. Little strokes fell great oaks. Creating CoRoLa, the Reference Corpus of Contemporary Romanian. *RRL*, LXIV(3):227-240.

CHAPTER 2

NATURAL LANGUAGE PROCESSING FOR E-LEARNING

EDUCATIONAL TEXT CONTENT GENERATION FOR ELEARNING SYSTEMS

ALEXANDR PARAHONCO¹, MIRCEA PETIC¹

¹*Vladimir Andrunachievici Institute of Mathematics and Computer Science,*

alexandr.parahonco@usarb.md

Alecu Russo Balti State University, Republic of Moldova

mircea.petic@math.md

Abstract

In recent years, a tremendous amount of information has been published on the internet. It has a negative impact on the field of eLearning, since teachers have to carry out a huge amount of preparatory work for the lessons while researching and surfing online. As part of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”, we created our web application for generating content that could solve this problem. At this point, we can find and collect information from the Internet. As a last step of our project implementation, we found an effective way of processing this information and providing our final product through automatic text summarization. This article discusses several approaches of summarization, including prompt engineering, abstractive-based summarization, and extraction-based summarization. A multilingual perspective is taken into consideration, covering the English, Romanian, and Russian languages. In addition, the information that served as the basis for conducting an experiment on automatic summary is analysed. Next, the study continues with the metrics used to evaluate the quality of the summarization result. Finally, the paper provides some thoughts regarding the results and approach to implement in the web application for the eLearning area.

Keywords: content generation, text summarization, data analysis, eLearning.

1 Introduction

Thanks to digital technological progress, the field of education becomes more and more diverse, and the importance of e-learning systems is continuously increasing (Ouatu and Gifu, 2020). Similarly, the possibilities offered by e-learning systems are becoming more and more varied. Even if digital technological wonders are available, the most important thing in an educational process is the content with

which an e-learning system is achieved. If in the initial stages of the emergence of e-learning systems they were filled-in by the teaching staff, the easing of this work by developing automatic content generation mechanisms would be a natural one. The desire to follow this thought even further seems welcome in the situation where the amount of information on the Web is continuously increasing in different languages. Considering the variety of content types possible for modern eLearning systems such as video, audio, graphic and textual, in this article we will focus on the automatic generation of textual content for e-courses.

E-courses are often known as online courses. These are the courses that have been delivered virtually as opposed to in person. Even yet, there are other ways that the courses can be delivered, including *live sessions*, *pre-recorded videos*, *PowerPoint presentations*, *PDFs*, *e-books*, and much more. The course materials are updated frequently and modified to meet the needs of the teaching process. Study resources have been enhanced throughout the Covid-19 epidemic with *lecture videos* and *several self-paced study examples* (Pobočiková *et al.*, 2023).

The aim of this article consists in researching the possibilities of automatic generation of educational content for e-course development and presenting our own vision together with the experience of developing our own mechanisms for automatically filling e-learning platforms with relevant content.

The paper is structured as follows. First, we describe the specifics of the evolution of eLearning from the beginnings until now. Below we have presented current examples where the elaborations have been created with reference to the automatic generation of content for eLearning. In the next section, our approach is presented with reference to the organization of the process of automatic completion of e-learning systems with educational content. In order to be able to solve the moments of identifying a large number of educational contents on a certain subject, the possibility of using available automatic summarizing tools was investigated. At the end of the study, we discuss the results obtained from the application of automatic summarization methods and formulate some conclusions following the research carried out in the article.

2 Modern eLearning Systems

Ever since the advent of personal computers, the idea of using them in education has been heavily researched. It seems the early stages of eLearning represented the possibility of recording educational content on external information supports of computer systems. The way to deliver educational content easily and quickly with minimal expenses has been the catalyst for major changes for today's eLearning-based education. This was a way to interact directly with the teacher outside of the class. We note that the prototype of the modern information support

was taken from the original information support, which was the book, the one that had to be read from the beginning to the end to complete a course. This is how the courses are built online now, with a well-defined scenario for the student to follow the steps in a timeline so that all the material predestined for an educational course is studied (Lowder, 2023).

The evolution led in 1995 to the emergence of Learning Management Systems (LMS) that have allowed the saving of large sums of money, by the help of participating in online courses because of the geographical distances of the students' place of residence. They allow not only the exchange of educational documents with students, but also student records, their attendance, test results, grades, and possibilities to issue notifications and tax receipts (Bouchrika, 2023).

The term *eLearning* emerged in 1999 at a lecture within the CBT Systems Company. The essence of the notion of *eLearning* consists in the use of digital means of any kind to obtain knowledge and skills in a certain field (Bouchrika, 2023).

Over time, it was passed from the simple acquisition of the material provided by the teacher to the more modern possibilities of computer systems such as interaction. The current trend is to deliver learning in simple ways and knowledge in dosed quantities so that it can be easily processed by students. Even so, with access to a multitude of online resources, full-time students want to obtain information that is current and relevant to future work as an employee in a chosen field. Beyond the quality of the content, it is important that the teacher knows the student's preferences for learning, the best ways, and that there are no limitations in possibilities for everyone alike (Lowder, 2023).

Other tools for eLearning are considered educational possibilities of social media platforms such as Facebook, YouTube, Twitter, and LinkedIn. These make education more and more accessible in the form of blogs, Twitter threads, short videos, and documentaries. At the same time, several notions related to eLearning emerged such as “mobile learning”, “gamification” and “social eLearning” which describe the variety of approaches in education. All this leads to the interest in eLearning not only from academic institutions but also from business ones, as eLearning has become a possibility for fast and effective training with limited resources (Bouchrika, 2023).

According to (Kumar *et al.*, 2022) the top 5 eLearning platforms used in academic institutions are: Moodle, Google Classroom, Blackboard, Top Hat and Canvas.

If we analyze what would be the most frequent and requested examples of practices used in eLearning, we could list the following (Kirstie, 2023): game-like quiz, microlearning, simple branching scenario, case study with polls, guided self-

reflection, scenario-based policy training, quick quiz, in-depth process training, in-depth learning: onboarding and compliance tests with question pools.

Currently the important notion is Knowledge Management System (KMS) in eLearning, thus we notice that the attention is transferred to current valuable knowledge. These include educational web platforms, learning management systems, corporate/public web portals, document management systems, scheduling systems, etc. practically everything that offers the possibility of carrying out educational processes at a distance (Softengi, 2023).

3 Examples of Applications for eLearning Content Generation

If we think about what the most time-consuming activity is related to the design of an eLearning course, we refer to the creation/updating of the educational content so that it would be current and original. It seems AI tools that generate on-demand educational content are designed around **keywords** that can suggest content that includes these entered keywords as a benchmark to generate the right content. *Rytr* is one such online tool that uses the notions of *artificial intelligence* based on the set language, content context and its use case along with keywords in a quick time is able to generate the right educational content (Jain, 2022).

The subject of automatic generation of educational contents is quite interestingly developed by the possibilities of artificial intelligence. In this regard, below there are several applications that appear to be useful in automatically generating educational content (Dane, 2023):

- **EdApp** – allows to generate micro lessons with the help of the Create AI tool, moreover it has the possibility of connection with Canvas for which there is a generation tool based on templates.
- **CourseBox** – has a built-in tool that quickly generates content using artificial intelligence, which based on a template can create a mobile-friendly course and/or transferred to any existing LMS.
- **NovoEd** – has the specifics of generating with the help of the latest generation artificial intelligence technologies that are based on learner data, thus generating personalized content.
- **MindSmith** – a tool that generates educational content assisted by artificial intelligence to also create lesson plans that can be integrated into existing LMS.
- **Courseau** – a technology that merges GPT3.5 and GPT4.0 chat to generate small courses based on a brief description.
- **CourseAI** – a content generator based on a simple topic. As a result, this application generates both possible titles and course contents, but also a course structure.

- **Mini Course Generator** - based on a concept, it will create the title, structure and content of the lesson, complemented with images and videos in such a way that they are attractive.
- **Lingio** offers mobile-friendly, gamified, and interactive courses available in more than 100 languages based on its own materials that are processed and possibly reformulated.
- **AdapteMy** suggests customized learning materials to address the areas where learning is needed.
- **Knewton**, apart from the possibility of generating personalized content based on artificial intelligence, it can also be integrated into our own LMS that we use.
- **Coassemble's AI Assistant** – able to generate educational content based on a title and short description (Warwick, 2023).

Following the research, it was established that overall, the generation of contents takes place especially in the case of the intention to provide the learner with personalized content.

4 Our Proposal for eLearning Content Generation

Taking into account everything presented in the previous sections, we will describe our approach to building the application for automatic generation of e-course content. In this regard, it should be mentioned that this solution could also be applied to students in perspective. In this case, the application plays the role of a chatbot, providing personalized content. In this case, the generation of the content goes through the following stages: determining the concept of the material, adapting the complexity of the content, analysing the characteristics of the learner, and developing the content based on the previously established ones (Gavrilović, *et al.*, 2017).

Web scraping is the fundamental technique in our web application. To find some information, we use the same query that we do with any search engine, like Yahoo, Bing, or Google. It is an ordinary text that the application employs to find the necessary resources using the Google search engine. The approach we offer goes through the following steps to achieve automatic completion of the e-course:

1. Network of synonymous words by means of a Web-crawler – we use the Wordsmyth¹ online dictionary to fetch synonyms for nouns from the search query in order to broaden the keywords area and simplify the user's search experience. This may result in a variety of search queries and bring qualitative material. For example, in the query

¹ Wordsmyth online dictionary - <https://www.wordsmyth.net>

“reforms of Peter the Great” the word “reform” has the synonym “improvement”. Thus, we can obtain the second query as “improvement of Peter the Great”;

2. Collection of links for generated requests in Google search engine and their storage in the database;
3. Gaining all the collected links from the database and extracting the necessary fragments by the URL from the Internet;
4. Generation of the processed information and its presentation to the users;
5. Content editing
 - a. The ability to view ready-made content (its original sources and move generated pieces).
 - b. Reformulation of the query for the new search with the ability to save some pieces.

The database plays an important role in this procedure. Being the primary storage of the application, it stores information about dictionaries that are used to find synonyms for each noun in the request. When the application searches for links with the targeted information, it addresses the database, first to create records with the links and, second, to update these records with the fetched information (Parahonco *et al.*, 2020).

At the third stage the application retrieves links from the database and fetches text, images, and video from within. A detailed view of this procedure can be demonstrated in Figure 1.

There are four steps in the system itself for every resource. The first step is to determine the type of resource. The function **getContenttype** retrieves the **URL** address' headers and returns the MIME type. Subsequently, the required method of managing resources is implemented. Here, various types of sources that represent uploaded materials on the Internet should be distinguished from a straightforward **web page**. As a result, the content of the webpage is immediately available, unlike resources like **doc**, **docx**, and **pdf**, which need to be downloaded and processed beforehand.

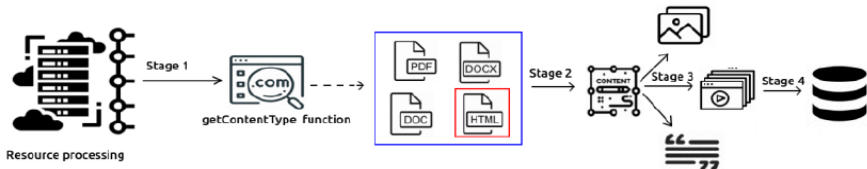


Figure 1. Content selection.

In terms of web page resources, stage 3 uses the content that was retrieved at stage 2 to search *image* and *video* addresses using their respective tags (**img** and **video**). The **findInText** function extracts meaningful portions of the text. After that, the database contains all of the acquired data.

Step 4 is responsible for the text processing and the extraction of the most valuable information. The big challenge is that we didn't know how to process all these sources (about 30) and provide them as one to the user. A literature review in this domain sheds light on the **summary approaches**. Summarization will help us interpret large amounts of the sources and present this resume as one final product, i.e., we can summarize one or many (in dependence on the summary approach) text sources and provide it to the user. We have determined three suitable approaches as prompt engineering, abstractive-based summarization, and extraction-based summarization. Each above-mentioned solution requires special steps of implementation in our application. We have already tested extraction-based summarization and published it. These results will be presented in the following section. In this article, we are going to test abstractive-based summarization in detail and compare the results with the previous ones.

While working with the application, it is essential for the user to change the generated content in different ways. In this respect we have prepared the editor with minimum requirements. As we did not work out the problem of content generation (i.e., summary approach), we could not go further. The editor is a set of toolbars for text formation, image, and video insertion.

The lower part of the interface allows the user to switch between fetched content. As can be noticed, along with the text, images and videos are also provided. If buttons are accessible, then the specific resources are found in the original content. These sources can be copied and inserted in the editor. The last one, in turn, can save the file as an HTML webpage (Parahonco and Petic, 2022).

5 Text Summarization as a Part of Content Generation for e-Courses

Automatic text summarization is a method that simplifies and adapts a source text to the needs of the user or task by extracting the most important information.

From a technological point of view, prompt engineering and abstractive-based summarization differ from extraction-based summarization. The first two technologies rely on neural networks and the same transformer framework, while extraction-based summarization relies on other fundamental NLP approaches. In order to shorten research time, we make the assumption that if abstractive-based summarization is a reliable approach, then prompt engineering is the same and should be investigated further.

5.1 Extraction-based Summarization

In our paper (Parahonco and Petic, 2023) we tried to solve three problems: evaluate text summaries, find the most effective approach and view the problem from a multilingual perspective.

Six texts from various sources with varying structures and topic domains were chosen in order to assess the quality of the extraction-based summarization (see Table 1).

Table 1. Descriptions of the selected datasets.

Nr.	Domain	Language	Chars	Words	Title
1	History	English	8599	1309	Reforms of Peter the Great
2	Geography	English	7939	1326	Pacific Ocean
3	Biology	Russian	10323	1356	Signs, characteristics of animal classification
4	Literature	Russian	53328	6879	Polyphony and dialogism; Bakhtin's Aesthetics
5	Informatics	Romanian	20579	3063	Artificial intelligence
6	Law	Romanian	11780	1652	Introduction to Civil Law

The following summary methods described in (Parahonco and Petic, 2023) were investigated: Luhn's heuristic Method, Edmundson heuristic Method, Latent semantic analysis (LSA), SumBasic algorithm, Kullback-Lieber (KL) Sum algorithm, Graph-based summarization (Reduction), LexRank algorithm, TextRank algorithm, Term Frequency method and Term Frequency-Inverse Document Frequency (TF-IDF). Literature review helped us to evaluate the summary by the following **metrics**: ROUGE (ROUGE-1, ROUGE-2, ROUGE-L), BLEU, METEOR and F-SCORE (calculated from ROUGE and BLEU).

However, as we concluded at the end of the article, it is necessary to include one more metric: expert opinion based on intrinsic evaluation. Comparison with the source document will help us measure how many main ideas of the source document are covered by the summary. Withal, this will help us to investigate such parameters as the readability and coherence of the shortened texts.

In this paper, we found the answers to all the questions. Thus, from a multilingual perspective, there are not big problems with Romanian and Russian texts while making the summary. Although tokenization and stopword lists seem to cause some obstacles, not all libraries were designed for one or another language. Therefore, it was obligatory to find the absent component and assemble it. As to evaluation metrics, the research has shown that we cannot be firmly confident in summary efficacy based on them. Finally, we have identified that the most important methods in text summarization are Luhn's heuristic method, TextRank, and the term frequency method (provided in decreasing order). Overall results can be found in Table 2, where we have counted the *maximum result* for

each method. It should be mentioned that summary length was arranged at 50% of the original texts for almost each approach (for exceptions TF (NLTK) and TF (spaCy) that apply threshold).

Table 2. Evaluation of the summary approaches.

Language/ Score	Rouge-1	Rouge-2	Rouge-L	Bleu	Meteor	F-1	F-2	F-L	Duration (sec)
English MAX	0.8042 Luhn	0.7580 Luhn	0.8042 Luhn	0.5322 Luhn	0.5895 TF-IDF (NLTK)	0.641 Luhn	0.625 Luhn	0.641 Luhn	2.8290 TF (spaCy)
English MIN	0.5045 TF (NLTK)	0.4147 TF (NLTK)	0.5045 TF (NLTK)	0.0601 TF (NLTK)	0.1252 TF (NLTK)	0.108 TF (NLTK)	0.105 TF (NLTK)	0.108 TF (NLTK)	0.0350 TF-IDF (Scikit Learn)
Romanian MAX	0.8875 TF (spaCy)	0.8221 TF (spaCy)	0.8875 TF (spaCy)	0.6857 TF (spaCy)	0.4973 TF-IDF (Scikit Learn)	0.774 TF (spaCy)	0.748 TF (spaCy)	0.774 TF (spaCy)	17.196 KL
Romanian MIN	0.4532 KL	0.3266 KL	0.4532 KL	0.0308 KL	0.1782 KL	0.058 KL	0.056 KL	0.058 KL	0.0389 TF-IDF (Scikit Learn)
Russian MAX	0.8598 TF (spaCy)	0.8037 TF (spaCy)	0.8598 TF (spaCy)	0.6315 TF (spaCy)	0.5243 TF-IDF (Scikit Learn)	0.728 TF (spaCy)	0.707 TF (spaCy)	0.728 TF (spaCy)	184.394 KL
Russian MIN	0.3336 TF-IDF (NLTK)	0.2389 TF-IDF (NLTK)	0.3336 TF-IDF (NLTK)	0.0026 TF-IDF (NLTK)	0.1376 TF (NLTK)	0.005 TF-IDF (NLTK)	0.005 TF-IDF (NLTK)	0.005 TF-IDF (NLTK)	0.0560 TF-IDF (Scikit Learn)

Scoring rubric was developed with the following criteria: storytelling value (in %), readability and coherence challenges (number of cases) and abrupt transition (number of cases).

Mainly, the scoring rubric includes *native ways* of content and text evaluation of intrinsic approach. The storytelling value was calculated from the proportion of stated topics and exposed ones. This criterion also included a list of important events or facts that were in the text.

The next parameter asserted readability and coherence by analysing storytelling sequences based on the original source. This criterion estimated the outlook of the sentences: first letter is capitalized, each sentence does not comprise any other symbols or characters that were not in the source and logical separation of text by paragraphs.

The last criterion estimated the smooth transition from one topic to the next one, or reference to the previously mentioned character, event, or fact.

According to this rubric, Table 3 lists the top four approaches from a multilingual perspective.

Table 3. Top four extraction-based summarization approaches.

Rank	Method	Percentage of quality
1	Graph-based summarization (Reduction)	81.25 %
2	Edmundson heuristic method	78.75 %
3	TextRank algorithm	77.5 %
4	Latent semantic analysis (LSA)	72.5 %
	Luhn’s heuristic method	

As it can be seen from the table, the application of the expert opinion in this paper has changed this list of the top extraction-based summary methods provided in the previous paper. The TextRank algorithm and Luhn’s heuristic method are still present in the list but have another rank, showing the dominance of TextRank over Luhn’s heuristic method.

5.2 Abstractive-based Summarization

The experiment was carried out with 13 LLMs for text summarization, including *bartLlargeCnn* (Lewis *et al.*, 2019), *finetunedBart* (Kabita Choudhary, 2023), *medical_summarisation* (Falcons.AI, 2023), *pegasus_xsum*, *pegasus_cnn_dailymail* (Zhang *et al.*, 2019), *t5_large*, *t5_small* (Raffel *et al.*, 2020), *mT5_m2o_russian_crossSum* (Hasan *et al.*, 2021a), *mbart_ru_sum_gazeta* (Gusev, 2020), *mT5_multilingual_XLSum* (Hasan *et al.*, 2021b), *rut5_base_summ* (Balobin, 2023), *ro_bart_512*, *ro_bart_large_512* (Taiatu, 2023). These models were selected by their high score in the rationing list on Hugging Face. The list contains *nine* models for the English language and *four* models both for the Romanian and Russian languages. Some models are multilingual, some bilingual and others monolingual.

In order to evaluate the quality of the abstractive-based summarization and compare these results with the extraction-based summarization, the same texts were selected from Table 1.

Among all the models, 3 of them can be distinguished: **BART**, **Pegasus** and **T5**. From the architecture point of view, we have BART and Pegasus **sequence-to-sequence models** and T5, an **encoder-decoder model** pre-trained on a multi-task mixture of unsupervised and supervised tasks. An interesting observation can be made regarding the **multilingual aspect** of these models. Generally, multilingual features are more characteristic of T5-like models; monolingual ones for BART and Pegasus-like models. Model estimation was performed due to the

upper mentioned metrics. This will allow us to compare both abstractive-based and extractive-based summarization. The obtained summaries let us highlight the following features: chaotic sequence of ideas, output oddities, specific interpretations, repetitions, input length, shifting the context of the summary, time consumption and summary length.

Input length and summary length can be the triggers for half of these pitfalls. Current LLMs can afford only 512 or 1024 tokens as a final product, regardless of the task, and are sensible for the provided number of tokens. On the one hand, we encountered many problems with the token number in the input text. It was either too large or too small for a specific model. Each of them has a different “understanding” about the token. Thus, it took time to handle this issue, though some warnings still persist. On the other hand, this volume of the summary is too small for the e-course development. Thus, instead of summarizing the whole text, we split it into parts and process them. As a result, we expand the summary, but obstacles occur. When we make a summary without the division of the text, for some models it is fluent and coherent. Thereby, the above-mentioned summary defects were observed under expanded summary versions. A chaotic sequence of ideas (topics) is essential for all the investigated models. This happens when topics become altered within the process of narration. This can worsen text comprehension and lead to false conclusions.

For example, the next passage about the reforms of Peter the Great (**medical_summarisation model**) “(1) **the government** reforms of Peter the Great aimed to modernize the Tsardom of Russia (later the Russian Empire) based on Western European models. (2) **despite Russia's vast size** and considerable natural resources, a number of factors, including corruption and inefficiency, hampered economic growth. (3) **the administration of the state**, which dominated most of Peter's reign, was strained to gather and manage resources. (4) **in 1708, Peter abolished** these old national subdivisions and established in their place eight governorates (guberniyas): Moscow, Ingermanland, Kiev, Smolensk, Archangelgorod, Kazan, Azov, and Siberian; this system divided the population unevenly and was extremely clumsy to manage. (5) **this system was** also significantly modified and revised under Peter 's rule.” **has five sentences and two topics:** the background and the administrative reform.

However, the narration of the text seems to mis-interpret the administrative reform (sentences 3, 4 and 5). According to the summary, it is not clear in the 4th sentence what “**these** old national subdivisions” refer to. As there was nothing spoken about it before.

The next sentence makes the situation even worse as it emphasizes the idea that **eight new governorates (guberniyas) were changed by the Petr**. In fact, the original text has a sentence above that explains the situation: “*The state was*

divided into uyezds, which mostly consisted of cities and their immediate surrounding areas; this system divided the population unevenly and was extremely clumsy to manage.” Then comes the 5th and the 4th sentences from the summary. Thus, becomes clear that old divisions in “uyezdz” were reformed into “eight governorates”.

Such kind of summary mis-interpretation due to chaotic sequence of ideas were seldom met. But the alternation of topics was presented in all the summaries. This, in turn, does not influence much text comprehension, though it depends on the case and the model.

The oddities of the output belong to different spellings, including incorrect way of starting the sentence and various symbols or chars in the sentence.

From the previous example, we can see that each sentence starts with a **lowercase letter**. This behaviour also preserves the **t5_small** and **t5_large** models. For example, a passage from the text about the Pacific Ocean “(1) *the Pacific Ocean* is a vast body of water teeming with activity. (2) *It covers up to 71% of the Earth’s surface and contains the deepest reaches known to humankind.* (3) *scientists* are just beginning to learn about life in the deepest parts of the oceans. This vast ocean basin is home to some of the most unique life forms on Earth. (4) *is a large island* of plastic, often impossible to see with the naked eye...” or, an example from Romanian summary about AI: “(1) *a, este cu adevarat inteligenta artificiala.* (2) *Ce este si cum va schimba viitorul.* (3) *de orice alta vietuitoare.* (4) *la tehnologii* de ultimă generatie, roboți capabili să înțeleagă și să decidă acțiunile care trebuie întreprinse și o lume futuristă în care mașinările și omul coexistă.” We would mention that the Romanian texts with diacritics for which the t5 model was applied in the result generated Romanian texts that contain some words without diacritics.

This type of text fault also affects **syntax rules**. Let’s consider the 4th sentence in the English example and the 1st sentence in the Romanian example. The sentences lack some speech parts to make up a coherent text.

The other type of mistake refers to the **symbols that suddenly appear in the text**. Let’s look at the summary about the reforms of Peter the Great provided by the **pegasus_cnn_dailymail** LLM: “Government reforms of Peter the Great aimed to modernize the Tsardom of Russia .<n>Reforms were intended to support the Great Northern War of 1700-1721 .<n>The establishment of the Table of Ranks was among the most audacious of Peter’s reforms .<n>Peter believed that targeted reform could not only strengthen his hold on power but increase the efficiency and thus better the lot of his people . <n>Reforms significantly changed the internal structure and administration of the state.”

The “<n>” symbol fills up the summary. Another instance can be found in the **ro_bart_large_512** LLM summary: „... au schimbat bazele de concepții si

conceptii. » » de dezvoltare » de schimbare » de schimbare». This passage is full of tiny symbols such as “»” and the line of dots “.....”.

rut5_base_summ model adds some **function words** like “**finished**”, “**Return**” and “**summarize**”: “В дальнейшем работы этих ученых проводились вплоть до XVI—XVII вв., когда было предпринято первое исследование более глубокой **систематизации** животного на основе его внутреннего строения, схожих с повадками и схожими повадочными повадки. **finished Return**, когда была предпринята первая попытка более глубокой **систематизации** природы на основе внутреннего строения животных, ...” and “Познакомьтесь с некоторыми видами хрящевых рыб и земноводных, которые обитают в **морей** и океане, где обитают хрящевые рыбы и морские хищники, которые, по данным **summarize**, являются вымершими панцерниками, которые царили в **морех**.”. The bold words, apart from the function words, are spelled wrong. It seems that this model distorts words throughout the text attempting to form coherent ideas.

Specifically for transformers and neural network models, the key concept is interpretation. As for the abstractive-based summarization, it is more than just rephrase the text; it produces new content. Even though, in recent years, technology has gained ground in this domain, there are questions about its training, fine tuning, number of parameters, and others.

Our experiment has demonstrated that texts were interpreted with some warnings. The first of them are **fake statements**. For instance, **bartLargeCnn** exposes the next formulation about the influence of Peter the Great: “*Peter also created the first Russian state bank, the Vostok Bank, which is still in existence today. He also created Russia's first post-secondary school, the Moscow State University, which still exists today. Peter is credited with the creation of the Russian Academy of Sciences, which was founded under his rule. He was also responsible for the establishment of the KGB, the successor agency to the **KGB**. He is also credited with creating the Kremlin, which remains in charge of most of Russia's intelligence services. Peter was also the father of the current Russian president, Vladimir Putin.*”.

This paragraph is full of *useful* and *awkward statements* that were not found in the original source. Let's examine each sentence in detail. The first one (*Peter also created the first Russian state bank, the Vostok Bank, which is still in existence today*) - is nonsense. Peter I never created the bank. Speaking about this sequence “*In 1860, the Emperor of Russia Alexander II signed a decree to establish the State Bank. This was the beginning of the history of the Bank of Russia*”. We should say that the second sentence is **half true**². Indeed, Peter I created the *School*

² History of the Bank of Russia - https://www.cbr.ru/eng/about_br/history/

of *Mathematical and Navigation Sciences* in 1701. But he never established the university in Moscow. It was in Saint-Petersburg in 1724 that was run under the Academy of Sciences³. This is emphasized in the next sentence: “*Peter is credited with the creation of the Russian Academy of Sciences ...*” The rest of the sentences have nothing to do with reality.

The same outcome has finetunedBart, pegasus_xsum, medical_summarization, mbart_ru_sum_gazeta and others. Generally, the sentences that were generated by the model play a negative role, rather than positive. Because they are not precise and can make false statements.

Following this topic, another interesting aspect can be observed in the paragraph about the Pacific Ocean: "Here, we discuss some of the key features of this great ocean and the issues affecting it. In this article, we review the history of the ocean basin and provide a comprehensive overview of the deepest reaches known to humankind." Here, we should pay attention to the third person in the narration, represented by "we". The style of the text changes from scientific to narrative, or something similar (Gifu *et al.*, 2016).

A shift in the summary context can be found in the **mbart_ru_sum_gazeta** LLM. Processing the text about animal classification, the model adds some cultural context: “*Российские ученые представили классификацию животных, которая объединяет пять основных царств, а в общей сложности ученым известно несколько миллионов видов, которые проживают или проживали на планете.*” Thus, the context of the text shifts from neutral to Russian, emphasizing **the role of Russian scientists** in animal classification procedures. Moreover, **it complements the idea with useful information** about the existence of several million species that live or have lived on the planet.

Another sample of the background change can be seen in the **mT5_m2o_russian_crossSum** large language model. The passage “*Ученые в США и Великобритании пришли к выводу, что относительно большинство видов являются домашними животными*” underlines the conclusions of USA and Great Britain scientists that relatively most species are pets. First of all, it is not the truth. Secondly, it adds the layer of national achievements to the text background. The same effect is present in the **mT5_multilingual_XLSum** model.

A curious process happens when in the text is mentioned the surname “Bakhtin” (referring to Mikhail Bakhtin) without the first name. All Russian models,

³ Реформы Петра I в области образования/Reforms of Peter I in the field of education - <https://old.bigenc.ru/text/5054236>

excepting **rut5_base_summ**, tried to guess the name with dozens of variations. Let's see some of them:

- “Архангел Михаил Бахтин” - Archangel Mikhail Bakhtin;
- “Евгений Бахтин” - Evgeny Bakhtin;
- “Стивен Бахтин” - Steven Bakhtin;
- “Ричард Бахтин” - Richard Bakhtin.

These interpretations complicate understanding the text, which is by itself too complex. Here it is necessary to refer to repetitions of the word sequences and repetitions of an idea that create the loop of inference. Working with the same text “Polyphony and dialogism; Bakhtin's Aesthetics”, such errors make it impossible to realize the meaning of the text.

Romanian models **ro_bart_512** and **ro_bart_large_512** provide the second type of repetitions. For instance, “Inteligența emoțională. Aceasta este și despre cum îți afectează capacitatea de a schimba caracteristicile la maturitate. O tehnologie artificială care combină caracteristicile de domenii diferite (sau cele mai complexe) pot fi folosite în moduri diferite, cum ar fi: caracteristicile sau caracteristicile și caracteristicile fundamentale sau de a și de a modifica caracteristicile unei tehnologii. Toate aceste caracteristici sunt revoluționizate și schimbătoare. Aceste caracteristici sunt folosite în mod convențional și sunt folosite pentru a schimba natura lucrurilor și a schimba realitatea. Sunt folosite în moduri specifice, dar și în scopuri specifice. și sunt utilizate în moduri diferite și identice cu cele specifice. Aceste tehnologii sunt inteligente și inteligente și complexe și inteligente, dar sunt inteligente și inteligente și identice ca și inteligente. și inteligente ca și ca și cum sunt evoluat în natura umană și inteligența și inteligența generală și inteligența artificială și inteligența artificială și inteligența umană ...”. The first underlined passage provides the loop of an idea of AI with the key word “characteristic” that changes interpretations and cannot stop. The second passage is more deterministic in this aspect.

Repetitions of the word sequences were observed in **pegasus_cnn_dailymail**, **pegasus_xsum** and **t5_large** LLMs. For example, “Peter believed that targeted reform could not only strengthen his hold on power but increase the efficiency and thus better the lot of his people .<n>Peter believed that targeted reform could not only strengthen his hold on power but increase the efficiency and thus better the lot of his people .” **or** “The resulting chain of volcanoes is roughly 25,000 miles long and springs to life where the Pacific tectonic plate slides against or collides into the other tectonic plates that circle it. <n>

Table 4. Evaluation of the summary approaches.

Score	Rouge-1	Rouge-2	Rouge-L	Bleu	Meteor	F-1	F-2	F-L	Duration (sec)	Lang.
<i>MAX</i>	0.5461 bartL largeCnn	0.4534 bartL largeCnn	0.5401 bartL largeCnn	0.4255 pegasus_ xsum	0.1150 bartL largeCnn	0.345 bartL largeCnn	0.324 bartL largeCnn	0.343 bartL largeCnn	951.15 pegasus_cnn_ dailymail	En
<i>MIN</i>	0 mT5_m2o_ russian_ crossSum	0 mT5_m2o_ russian_ crossSum	0 mT5_m2o_ russian_ crossSum	0 mT5_multi lingual_ XLSum	0 mT5_multi lingual_ XLSum	0 mT5_multi lingual_ XLSum	0 mT5_multi lingual_ XLSum	0 mT5_multi lingual_ XLSum	11.399 t5_small	
<i>MAX</i>	0.3681 t5_small	0.2696 t5_small	0.3643 t5_small	0.0159 t5_large	0.0282 t5_small	0.03 t5_large	0.029 t5_large	0.03 t5_large	1417.9 t5_large	Ro
<i>MIN</i>	0.3011 t5_large	0.1432 t5_large	0.2942 t5_large	0.0060 t5_large	0.0047 t5_large	0.012 t5_large	0.012 t5_large	0.012 t5_large	24.355 t5_small	
<i>MAX</i>	0.6899 mbart_ru_ sum_gazeta	0.5714 mbart_ru_ sum_gazeta	0.6899 mbart_ru_ sum_gazeta	0.6370 rut5_base_ summ	0.3035 mbart_ru_ sum_gazeta	0.588 rut5_base_ summ	0.499 mbart_ru_ sum_gazeta	0.567 rut5_base_ summ	3739.9 mT5_m2o_ russian_ crossSum	Ru
<i>MIN</i>	0.1197 mT5_m2o_ russian_ crossSum	0.0142 mT5_m2o_ russian_ crossSum	0.1118 mT5_m2o_ russian_ crossSum	0.005 mT5_multi lingual_ XLSum	0 mT5_multi lingual_ XLSum	0.01 mT5_multi lingual_ XLSum	0.009 mT5_multi lingual_ XLSum	0.01 mT5_multi lingual_ XLSum	187.39 mT5_multi lingual_ XLSum	

The resulting chain of volcanoes is roughly 25,000 miles long and springs to life where the Pacific tectonic plate slides against or collides into the other tectonic plates that circle it.”

Finally, time consumption is not the last parameter to consider. It should be taken into account that the high value of time does not stand for high quality of results and inverse. The shortest time is 11 seconds, and the slowest is 3739 seconds. The average time is about 696 seconds, and the median is 271 seconds. Results can be found in Table 4, where we have counted the *maximum result* for each method.

6 Discussions on the Results

The undertaken analysis of the abstractive-based summarization approach, along with the expert review (the same scoring rubric), allowed us to conclude that it is less effective than extractive text summarization approaches.

Basically, all the LLMs committed lots of errors that worsened perceptions of the information. Their summaries provided overly compressed content and specific interpretations. Hence, the e-course content could be poorly developed or not well thought out. Nevertheless, such models as bartLlargeCnn, finetunedBart and pegasus_cnn_dailymail had few errors and pretty good summary. Though their time consumption is too much (112, 109, and 860 seconds, respectively). It is from 2 to 16 minutes. Compared to 1.99 seconds for reduction, 0.2 seconds for the Edmundson heuristic method, and 1.96 seconds for the TextRank algorithm.

From a multilingual point of view, there are problems in finding suitable LLM. For Romanian and Russian languages models are poorly developed. These models

are not yet ready for production release. In contrast with extractive text summarization approaches that do not have any problems with language-specific texts (excepting tokenization and stopword lists).

Accordingly, to fulfil the aim of the project to create the application for e-course content generation, extractive text summarization approaches should be applied. The best of them is graph-based summarization (reduction), the Edmundson heuristic method, and the TextRank algorithm. Unfortunately, transformers and neural network models do not satisfy any requirements of the project. These models use many resources and provide little quality now.

7 Conclusions

Thanks to digital technological progress, the field of education is becoming more and more diverse, and the importance of eLearning systems is continuously increasing. The possibilities offered by eLearning systems are also becoming more and more varied. Considering the variety of content types possible for modern e-learning systems, such as video, audio, graphic, and textual, in this article we focused on the automatic generation of textual content. Within the scope of this article, several approaches have been examined, including prompt engineering, abstractive-based summarization, and extraction-based summarization. A multilingual perspective was taken into consideration, namely the English, Romanian, and Russian languages. We have discovered that Romanian and Russian language models are poorly developed and cannot be used in real projects.

The experiment has demonstrated that LLMs suffer from different problems that hamper text understanding. From the long list should be emphasized the chaotic sequence of ideas, specific interpretations, repetitions of different kinds and shifting the context of the summary. They could make text comprehension even more impossible and lead to false conclusions.

The time consumption component proves that abstractive-based summary approaches take a long time, in contrast with extractive-based summary ones. Thus, time consumption is 112, 109, and 860 seconds, respectively, for bartLlargeCnn, finetunedBart, and pegasus_cnn_dailymail LLMs. Compared to 1.99 seconds for reduction, 0.2 seconds for the Edmundson heuristic method, and 1.96 seconds for the TextRank algorithm,

Finally, extractive text summarization techniques should be used to achieve the project's goal of developing an application for the creation of e-course content. Graph-based summarization (reduction), the Edmundson heuristic method, and the TextRank algorithm are the finest among them. Unfortunately, neither

transformers nor neural network models meet any of the project's requirements. These models currently offer poor quality and consume a lot of resources.

Acknowledgements

This article was written within the framework of the research project 20.80009.5007.22 “Intelligent information systems for solving ill-structured problems, processing knowledge and big data”.

References

- Balobin, D. 2023. Finetuned AI-Forever/ruT5-Base for Text and Dialogue Summarization [online]. [Accessed 10.11.2023]. Access mode: <https://huggingface.co/d0rj/rut5-base-summ>
- Bouchrika, I. 2023. History of eLearning: Evolution from Stenography to Modern LMS Platforms. In: Research.com, 2023. [online]. [Accessed 09.11.2023]. Access mode: <https://research.com/education/history-of-elearning>
- Dane, D. 2023. 10 AI Course Creators for 2023. EdApp. September 15. <https://www.edapp.com/blog/ai-course-creator>
- FalconsAI. Model Card: T5 Large for Medical Text Summarization [online]. [Accessed 10.11.2023]. Access mode: https://huggingface.co/Falconsai/medical_summarization
- Gavrilović, N., Jovanović, S., and Mishra, A. 2017. Personalized Learning System Based on Student Behavior and Learning Style. The 8th International Conference on eLearning (eLearning-2017), 28 - 29 September 2017, Belgrade, Serbia, pages 34 - 39. ISBN 978-86-89755-13-8
- Gifu, D., Dascălu, M., Trăușan-Matu, Ș., and Allen, L K. 2016. Time Evolution of Writing Styles in Romanian Language. *Proceedings of the International Conference on Tools with Artificial Intelligence*, Bourbakis, N., Esposito, A., Mali, A. *et al.* (eds.), pages 1048-1054
- Greany, K. 2023. Top 10 Best elearning Examples to Inspire You in 2023. Elucidat. June 28. <https://www.elucidat.com/blog/best-elearning-examples>
- Gusev, I. 2020. Dataset for Automatic Summarization of Russian News. In A. Filchenkov, J. Kauttonen, and L. Pivovarov (Eds.), *Artificial Intelligence and Natural Language*. Springer International Publishing (pages 122-134).
- Hasan, T., Bhattacharjee, A., Ahmad, W.U., Li, Y.-F., Bin Yong-Kang, and Shahriyar, R. 2021a. CrossSum: Beyond English-Centric Cross-Lingual Abstractive Text Summarization for 1500+ Language Pairs. *CoRR*, *abs/2112.08804*. <https://arxiv.org/abs/2112.08804>
- Hasan, T., Bhattacharjee, A., Islam, Md. S., Mubasshir, K., Li, Y.-F., Kang, Y.-B., Rahman, M. S., and Shahriyar, R. 2021b. XL-Sum: Large-Scale Multilingual Abstractive Summarization for 44 Languages. *Findings of the Association for Computational Linguistics: ACL-IJCNLP*. 4693-4703. <https://aclanthology.org/2021.findings-acl.413>
- Jain, A. 2022. 4 Ways Artificial Intelligence is Transforming the eLearning Industry. Commlab India. August 18. <https://blog.commlabindia.com/elearning-design/ai-ways-transform-elearning-industry>

- Choudhary, K. 2023. Fine Tuned BART for Conversation Summary [online]. [Accessed 10.11.2023]. Access mode: <https://huggingface.co/kabita-choudhary/finetuned-bart-for-conversation-summary>
- Anal Kumar, M. S., Sharma, V., Ashika, A., Ali, A. S., & Narayan, A. 2022. E-Learning Platforms Analysis. *Modern Review: Hong Kong Journal of Social Sciences*, Vol. 60: 921-935. DOI: <https://doi.org/10.55463/hkjss.issn.1021-3619.60.89>
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *CoRR*, *abs/1910.13461*. <http://arxiv.org/abs/1910.13461>
- Lowder, J. 2023. The Evolution of eLearning for the Modern Learner. Performance Matters Blog. <https://www.gpstrategies.com/blog/the-evolution-of-elearning-for-the-modern-learner>
- Ouatu, B. and Gifu, D. 2020. Chatbot, the Future of Learning? *Proceedings of the 5th International Conference on Smart Learning Ecosystems and Regional Development (SLERD 2020)*, in Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education, Springer, pages 263-268.
- Parahonco A., Petic M., and Negara C. 2020. The Model of Web Crawler for Expansion the Scope of Initial Search. In: *Workshop on Intelligent Information Systems: Proceedings WIIS2020*. 04-05 December 2020, Chisinau. 139-150. ISBN 978-9975-68-415-6.
- Parahonco A. and Petic M. 2022. Elearning Content Processing Situations and Their Solutions. In: *Proceedings of Workshop on Intelligent Information Systems: WIIS2022*. October 06-08, 2022, Chisinau. 154-159. ISBN 978-9975-68-461-3.
- Parahonco A. and Petic M. 2023. Different Aspects on Extractive Text Summarization as a Part of Content Generation for e-Courses. In: *Proceedings of Workshop on Intelligent Information Systems: WIIS2023*. October 19-21, 2023, Chisinau. 169-180. ISBN 978-9975-68-492-7.
- Pobočíková, I., Sedláčková, Z., Michalková, M., and Malacká, Z. 2023. E-learning as an Integral Part of Study - Experience with the Use of e-Learning Course in the Subject Numerical Methods and Statistics. *Edulearn23 Proceedings*, 3964-3969. <https://doi.org/10.21125/edulearn.2023.1067>
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(140), 1-67.
- Softengi. 2023. 5 E-learning Trends: How Should a Modern LMS Look Like. [online]. [Accessed 17.11.2023]. Access mode: <https://softengi.com/blog/5-e-learning-trends-how-should-a-modern-lms-look-like>
- Taiatu, I. 2023. A Pre-trained from Scratch BART Base Model (140M parameters) [online]. [Accessed 10.11.2023]. Access mode: <https://huggingface.co/Iulian277/ro-bart-512>, <https://huggingface.co/Iulian277/ro-bart-1024>
- Warwick, Kelsy. 2023. Learning Management System and eLearning Blog. Coassemble. How to Create Engaging Content in Seconds with Coassemble's AI Assistant., March 2, 2023 [online]. [Accessed 15.11.2023]. Access mode:

<https://coassemble.com/blog/how-to-create-engaging-content-in-seconds-with-coassembles-ai-assistant>

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *CoRR*, *abs/1912.08777*. <http://arxiv.org/abs/1912.08777>

METAPHOR MASTERY: AUGMENTED REALITY DICTIONARY FLASHCARDS FOR E-LEARNING

OLESEA CAFTANATOV¹, LUDMILA MALAHOV¹,
INGA ȚIȚCHIEV¹, DAN TĂLĂMBUȚĂ¹,
DANIELA CAGANOVSKI¹, EUGENIA MINCU²

¹*Vladimir Andrunachievici Institute of Mathematics and Computer Science,
USM MD2028, Chișinău, Academy 5 str., RM*

{olesea.caftanatov, ludmila.malahov, inga.titchiev}
@math.md

{dantalambuta, dana.caganovschi}@gmail.com

²*Bogdan Petriceicu-Hasdeu Institute of Romanian Filology, USM
MD2028, Chișinău, Academy 3/2 str., RM*
jana_mincu@yahoo.fr

Abstract

This article explores the development and implementation of Augmented Reality (AR) dictionary flashcards for e-learning to enhance metaphor mastery. The study of metaphor terms is pertinent to the study of language and human communication, including specialized communication. The design of augmented flashcards for metaphor mastery involves the creation of physical flashcards and the use of AR applications. By combining digital enhancements with physical cards, learners can access additional information, including videos, illustrations, and etymological details, enriching their understanding of metaphorical terms.

Keywords: Augmented Reality, Flashcards, Metaphor terms, Educational Technology.

1 Introduction

The research presents an innovative and interactive educational tool that combines traditional flashcards with augmented reality technology. These augmented flashcards are designed to enhance the learning experience by incorporating digital elements, typically through the use of smartphones, tablets, or other AR-enabled devices. The focus of the learning process is on metaphorical terms. Metaphor terms are words or expressions used to describe objects or concepts figuratively or symbolically beyond their literal meanings, especially within specialized communication situations.

Metaphors are commonly used in language to make descriptions more vivid and to convey abstract concepts in more relatable terms. They are a powerful tool in literature, poetry, and everyday communication, allowing people to create imagery, draw parallels, and provide insight into various subjects and ideas. For example, "Time is money" is a metaphor that implies time, like money, should be used wisely and not wasted.

For this research, we utilized the metaphor terms from the book "Teaching Metaphor Mini Romanian-English Dictionary of Metaphor Terms with Illustrations and Etymologies" by Mincu *et al.* (2023). The dictionary is intended for the educational environment – teachers, students, pupils, translation specialists, experts in various fields, etc. – but also for uninformed readers interested in the essence and value of metaphor for universal human existence, as it arises from culture and reflects beliefs and experiences.

2 The Power of Metaphors

Metaphor is studied as a philosophical concept, a linguistic phenomenon, and as a cognitive means of understanding. In philosophy, the founder of the doctrine of metaphor is considered to be Aristotle, who defines metaphor as an inappropriate name transferred from the genus to the species, or from the species to the genus, or from species to species, or by analogy (Aristotle, 1984). A metaphor is a hidden comparison, achieved by applying the name of one object to another, thus revealing some important features of the second.

Metaphors are a fundamental linguistic and cognitive concept used in language and communication to convey meaning and express abstract ideas by drawing parallels between different concepts or domains. They involve describing something in terms of something else to provide insight, create imagery, or make complex ideas more accessible. Metaphors are pervasive in everyday language and play a crucial role in shaping how we think, understand, and communicate about the world around us. Metaphors play a significant role in language, culture, and everyday communication by enriching expression, conveying complex ideas, and shaping the way we perceive and understand the world. Here are some key aspects of their role:

- **Enhancing Expression:** Metaphors provide a powerful means of enriching language. They add depth and imagery to communication, making it more engaging and memorable. For example, saying "Inima de piatră (heart of stone)" - This metaphor is used to describe a person who appears to be devoid of empathy or compassion.

- **Conveying Abstract Ideas:** Metaphors help convey abstract, intangible, or complex concepts by relating them to something more concrete or familiar. For instance, using the metaphor "A zbura cu gândul (fly with your thoughts)" – this metaphor refers to the ability to imagine or dream without restrictions, resembling the flight of thoughts in the world of imagination.
- **Cultural Significance:** Metaphors are deeply embedded in a culture's language and thought patterns. To fully grasp a language and its nuances, one must comprehend the cultural metaphors that shape it. This is especially important for language learners seeking to integrate into a new culture.

The research (Mincu *et al.*, 2023) explores the didactics of metaphor, focusing on metaphorical terms used in specialized languages. The work introduces a Romanian-English dictionary of approximately 200 metaphorical terms, providing a conceptual classification, etymologies, and illustrations. In the initial phase of our research, we have successfully developed several augmented flashcard models to serve as prototypes. These models encompass **mythological metaphors, sacral metaphors, anthropomorphic metaphors, chromatophore metaphors, and artifact metaphors**. While some models convey a singular meaning, others exhibit two or more meanings, prompting the creation of flashcards tailored to each interpretation. Each flashcard is enriched with multimedia illustrations, including videos, GIFs, and images that visually represent those metaphorical terms.

To enhance the educational value of the flashcards, we've included detailed information on the backside, offering insights into the origin and etymology of each term. This additional context aims to provide users with a holistic understanding of the metaphor's historical and linguistic roots. Notably, the introduction of the term "terminological metaphor" stands out as an innovative contribution, offering a fresh perspective on interlinking terms from different domains.

This conceptual framework deepens our comprehension of the metaphorical creation process and its practical application in specialized languages. As we progress, it's important to note that the manual creation of flashcards, along with the intricate process of associating meanings with corresponding metaphors, is currently underway. This hands-on approach allows for meticulous attention to detail, ensuring that each flashcard effectively captures the essence of its metaphorical term.

3 Challenges in Learning Metaphors

Learning metaphors poses several challenges, especially for individuals whose first language is not English. According to (Littlemore, 2011) a research study focused on the comprehension of metaphors among international students whose first language is not English, attending undergraduate lectures at a British university. The study involved participants identifying challenging words or multiword expressions in extracts from academic lectures and subsequently interpreting metaphors within those extracts.

The research aimed to determine the proportion of metaphorical items among reported difficulties and to assess the participants' awareness of their metaphor interpretation challenges. The findings revealed that among the items identified as challenging, approximately 40% involved metaphorical expressions, even when composed of words familiar to the students. Additionally, when international students misinterpreted metaphors, they demonstrated awareness of their difficulty in only about 4% of cases. Given the significant role metaphors play in university lectures for functions such as explanation and evaluation, it suggests that these students might be missing valuable learning opportunities.

The challenges in learning metaphors extend beyond English and are applicable to languages like Romanian as well. Similar difficulties arise due to the figurative nature of metaphors, cultural nuances, and the idiosyncrasies inherent in the use of metaphorical expressions. Here's how these challenges may manifest in the context of Romanian:

- **Cultural Variations:** Metaphors are often deeply rooted in cultural contexts, and learners from different cultural backgrounds may find it challenging to grasp the intended meaning behind metaphors that are culturally specific. Metaphors in Romanian, like in any language, often draw upon cultural references and context. Learners with different cultural backgrounds may encounter challenges in grasping the intended meanings if they are not familiar with these cultural nuances.
- **Idiomatic Nature:** Metaphors are inherently idiomatic, meaning that their meanings are not always deducible from the literal meanings of the words used. This can be confusing for learners who are accustomed to interpreting language more literally.
- **Figurative vs. Literal Understanding:** Distinguishing between figurative and literal language is crucial in understanding metaphors. Learners may struggle to identify when a phrase or expression is meant to be taken metaphorically rather than literally.

- **Limited Exposure:** Metaphors often rely on shared cultural experiences, and learners with limited exposure to the language may not have encountered certain metaphors in everyday life, making them harder to understand.
- **Lack of Awareness:** Learners may not always be aware of their difficulties in interpreting metaphors. This lack of awareness can hinder the learning process, as addressing challenges requires recognition of the areas that need improvement.
- **Fear of Misinterpretation:** Similar to English learners, those acquiring Romanian may have concerns about misinterpreting metaphors, potentially leading to hesitancy in language use.

Addressing these challenges requires a combination of targeted language instruction, exposure to diverse linguistic contexts, and opportunities for learners to practice and engage with metaphors in meaningful ways. In this context, we decided to develop an augmented application for learning metaphorical terms.

4 Augmented Reality in e-Learning

The integration of Augmented Reality (AR) into e-learning has revolutionized the educational landscape, offering a paradigm shift in how learners engage with digital content. Augmented Reality introduces a new era of interactive learning, surpassing traditional methods by allowing real-time engagement with digital content, (Țițchiu *et al.*, 2023). This dynamic approach captures and sustains learners' attention in ways previously unattainable. Learning metaphors through augmented reality presents a unique set of challenges and opportunities that contribute to the complexity of this educational endeavour.

Augmented reality provides a sensory-rich environment, but translating abstract metaphorical concepts into immersive experiences requires careful design. Creating tangible, sensory-rich representations for metaphors challenges educators to balance realism and metaphorical essence. In the initial iteration of our application, we employed the visualization of metaphoric terms. An illustrative example will be presented in the following section.

5 Designing Augmented Metaphor Mastery Flashcards

Flashcards are a learning and study tool commonly used to aid in memorization and quick review of information (Senzaki, 2017). They consist of a set of small cards, typically made of paper. The Augmented flashcards are digital equivalents that can be accessed by augmented application. Each card usually contains a question, term, or concept on one side and its corresponding answer, definition,

or explanation on the other side. In our case, augmented flashcards are designed to learn the metaphorical terms. In general, augmented flashcards of metaphor aims to teach and understand metaphors in a structured and guided manner, nurturing communication skills and the interpretation of metaphorical meanings in various contexts, both common and specialized. Here's how augmented reality flashcards work:

Step 1. Designing Physical Flashcards. Like traditional flashcards, AR flashcards are physical cards with information, such as words, images, or symbols, printed on them. Two examples of physical flashcards are presented in Figure 1. A physical flashcard has quite a few tasks to accomplish. Besides the fact that it has to capture student's attention, entice them to pick up their mobile and scan the image, it should have a high quality to let the AR experience come to life. When it comes to designing physical flashcards, there are few recommendations on obtaining the best performance from physical target images.

Physical flashcards should be rigid, not flexible. A hard material such as card stock or plastic is better than a simple printed piece of paper, because the flexibility of the printed piece of paper can make it difficult for the object to stay in focus. However, paper flashcards are easily reproducible and widely available, so make sure to fix them to a non-flexible surface. Additional insights into design requirements are explored in the study conducted by (Caftanatov *et al.*, 2022).

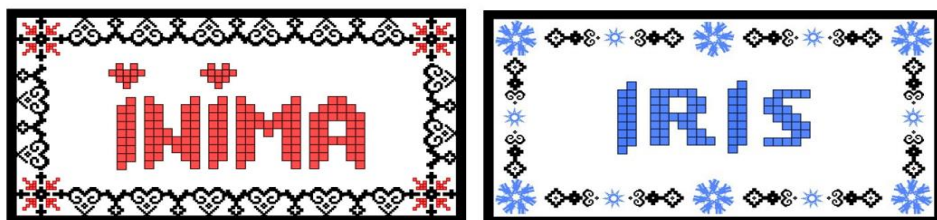


Figure 1. Physical Flashcard with metaphor “*inima*” that means *heart* on left side and metaphor “*iris*” with the same meaning in English on the right side.

Step 2. Augmented Reality App: Users need to have an AR app installed on their device, such as a smartphone or tablet.

Step 3. Scanning: When the user points their device's camera at the flashcard, the app recognizes and scans the card.

Step 3. Digital Enhancements: After scanning, the app overlays digital content on the physical flashcard, including supplementary details, illustrations, and etymological information about the metaphorical term. The flashcard represented in Figure 2 showcases the digital enhancements of augmented flashcards. Each flashcard comprises two sides, namely the front and back. On the

front side, there is video content pertaining to "heart" and "iris" respectively. On the back side of the flashcards, users can access the etymological content of the metaphoric term.



Figure 2. Example of front and back side of augmented flashcards for metaphor *inima* and *iris*.

6 Future Prospects and Challenges

Looking ahead, the evolution of augmented flashcards related to metaphor terms holds both promising prospects and challenges. Augmented flashcards may integrate seamlessly with language learning applications, providing users with a cohesive and comprehensive language acquisition experience. This integration could involve real-time language translation, pronunciation guidance, and cultural context insights. One of the future prospects we anticipate is an increase in interactivity based on augmented flashcards evolving to offer more engaging and immersive experiences. This may involve incorporating advanced technologies like gesture control or voice recognition.

Developing an augmented flashcard for illustrating metaphor terms is a big challenge. For instance, even **symbolic visualization** could give us a headache. Metaphors often rely on symbolism, and representing these symbols visually in augmented reality demands thoughtful consideration. Deciphering the symbolic visual cues within the augmented space adds a layer of complexity to the learning process.

For example, consider the metaphorical expression **"time is a thief"**. In this metaphor, time is compared to a thief, suggesting that it steals moments from our lives. The challenge in understanding this metaphor lies in recognizing that time is not literally a thief but that the comparison is made to convey the idea of time slipping away or robbing us of opportunities. The illustration of an hourglass with disappearing sand or a clock with hands resembling thief tools could visually

represent this metaphorical concept. Such examples help learners grasp the figurative nature of metaphors and the nuanced meanings they convey.

Another challenge could be **resource integration**. Augmented reality draws on various digital resources to create metaphorical experiences. Integrating diverse resources seamlessly and ensuring their alignment with educational objectives lead to challenges in curriculum design and technological implementation.

Technical Accessibility could also be a challenge. Ensuring technical accessibility for all learners, regardless of their technological proficiency, remains a challenge. Augmented reality tools should be user-friendly and accessible to a broad audience, overcoming potential barriers to entry. In navigating these challenges, educators have the opportunity to leverage augmented reality as a powerful tool for enhancing metaphorical understanding. Thoughtful design, cultural sensitivity, and a focus on diverse learning styles can contribute to the successful integration of metaphors into augmented reality learning experiences.

7 Conclusion

In conclusion, this research introduces an innovative approach to language learning through the integration of augmented reality technology with traditional flashcards. The focus of this educational tool is on metaphorical terms, exploring the profound role metaphors play in language, communication, and cultural expression. As mentioned in article, our research, currently in its initial phase, has yielded the development of augmented flashcard models focused on various metaphorical categories. These categories include mythological metaphors, sacral metaphors, anthropomorphic metaphors, chromatophore metaphors, and artifact metaphors. While some models convey singular meanings, others exhibit multiple interpretations, prompting the creation of dedicated flashcards for each nuanced understanding.

The research then delves into the realm of augmented reality in e-learning, emphasizing its transformative impact on education. Augmented flashcards designed for metaphor mastery represent a dynamic shift in how learners engage with abstract concepts. The challenges in this integration include balancing realism and metaphorical essence, ensuring content quality, and addressing technological accessibility.

The flashcards are designed to be comprehensive learning tools, each equipped with multimedia illustrations such as videos, GIFs, and images to vividly represent the metaphorical terms. On the reverse side of each flashcard, we've incorporated additional layers of information, providing users with insights into the origin and etymology of each term. This multifaceted approach ensures a holistic learning experience, combining visual representation with linguistic context.

References

- Mincu, E., Macovei, D., Rotari, N., and Negru, C. 2023. *Didactica Metaforei Mic Dictionar Român-Englez de termeni-metaforă*. Universitatea de Stat din Moldova, Institutul de Filologie Română "Bogdan Petriceicu Hasdeu". Chișinău: [Epigraf], 2023 (F.E.-P. "Tipografia Centrală"). 328 pages. ISBN 978-9975-60-521-2.
- Aristotle. 1984. *Сочинения в четырех томах*. (Translated title: *Essay in four volumes*). T4. Science Academy. Institute of Filology. Ed. "Мысль". Moscova. <https://pavroz.ru/files/aristotle4.pdf>
- Littlemore, J., Chen, P. T., Koester, A., and Barmden, J. 2011. Difficulties in Metaphor Comprehension Faced by International Students whose First Language is not English. *Applied Linguistics*. 2011:1-23. DOI:10.1093/applin/amr009
- Țițchiev, I., Caftanatov O., Tălămbuță, D., Iamandi, V., and Caganovschi, D. 2023. An Approach to Augmented Reality Classification and an Example of Its Usage for Application Development with VAK Learning Style Markers. In: *Computer Science Journal of Moldova*, vol. 31, 2(92):248-271.
- Senzaki, S., Hackathorn, J., Appleby, D. C., and Gurung, R. A. 2017. Reinventing Flashcards to Increase Student Learning. *Psychology Learning & Teaching*, 16(3):353-368. <https://doi.org/10.1177/1475725717719771>
- Caftanatov, O., Titchiev, I., Iamandi, V., Talambuta, D., and Caganovschi, D. 2022. Developing Augmented Artifacts Based on Learning Style Approach. *Proceedings of the Workshop on Intelligent Information Systems WIIS2022*, October 06-08, 2022, Chisinau, Republic of Moldova, 89-103, ISBN 978-9975-68-461-3

ENHANCING LANGUAGE LEARNING THROUGH AUGMENTED REALITY: A FOCUS ON MULTIWORD EXPRESSIONS

OLESEA CAFTANATOV¹, LUDMILA MALAHOV¹,
INGA ȚIȚCHIEV¹, DAN TĂLĂMBUȚĂ¹, TUDOR BUMBU¹,
ANA NASTASIU¹, VICTORIA BOBICEV²

¹*Vladimir Andrunachievici Institute of Mathematics and Computer Science,
USM MD2028, Chișinău, Academy 5 str., RM*

olesea.caftanatov@math.md

{ludmila.malahov, inga.titchiev, tudor.bumbu}@math.md

{dantalambuta, ana.nastasiu19}@gmail.com

²*Technical University of Moldova*

MD2004, Chișinău, Ștefan cel Mare bd. 168 str., RM

victoria.bobicev@ia.utm.md

Abstract

This research focuses on the use of Augmented Reality (AR) as an innovative tool to enhance the comprehension and application of Multiword Expressions (MWEs) in the Romanian language. Furthermore, as a member of Cost Action CA21167 (Universality, diversity, and idiosyncrasy in language technology), we are in the process of compiling a collection of Romanian Multiword Expressions (MWEs) that share similar meanings with their English counterparts. This initiative aims to facilitate collaboration with our colleagues from Albania and Spain, fostering a cross-cultural exchange of MWEs and promoting a deeper understanding of linguistic universality, diversity, and idiosyncrasies in language technology.

Keywords: Augmented Reality, Language Learning, Multiword Expressions, Romanian Language, Educational Technology.

1 Introduction

This research explores the synergistic potential of augmented reality (AR) technology in the context of language learning, with a specific focus on Multiword Expressions (MWEs). Multiword Expressions are linguistic constructs that often defy literal translation and represent the idiosyncrasies of a language. MWEs are

an integral part of everyday language and are frequently used in both written and spoken communication. Failing to understand or use MWEs properly can lead to misunderstandings or awkward communication. We have compiled a corpus comprising 17,200 Romanian multiword expressions along with their inflected forms sourced from the internet. Among these, there are 1,500 unique expressions. Both datasets, one consisting of multiword expressions with inflection forms and the other with unique forms, lack contextual information. Additionally, these collections exclusively feature Moldovan Romanian expressions.

At this initial stage of our research, we have successfully amassed a comprehensive dataset of multiword expressions (MWEs). However, a crucial aspect that remains in progress is the acquisition of detailed descriptions for each unique multiword expression. To elucidate, our dataset encompasses 25 expressions associated with "**ceas (clock)**", 57 expressions linked to "**vreme (time)**", 17 expressions connected to " **timp (time)**", and 2 expressions related to "**oră (hour)**". However, in our application, we presently possess Romanian descriptions for only 14 expressions, with the remaining expressions still undergoing the description collection process.

The descriptions assigned to these multiword expressions are meticulously curated from diverse sources, including internet references, Dex Online, the e-book "*Dictionar de expresii romanesti in contexte*" by (Ilinca, 2015), and insights derived from the collective experience of our authors. It is essential to emphasize that each description is being manually gathered to ensure accuracy and relevance.

2 Understanding Multiword Expressions

A multiword expression is a linguistic term used to describe a group of words that commonly appear together and function as a single unit of meaning. These expressions often consist of multiple words that, when combined, convey a specific and non-literal meaning that may not be easily deduced from the individual words. MWEs are also sometimes referred to as multiword units, multiword sequences, or collocations. Multiword Expressions are intricate linguistic constructs that often present challenges in language acquisition due to their non-literal and idiosyncratic nature (Espinal and Mateu, 2019). These expressions, including idioms, collocations, and phrasal verbs, are vital for fluency in any language. The research team compiled an extensive corpus of 1500 unique Romanian MWEs from a variety of sources, highlighting their importance in language proficiency and communication (Vercellotti *et al.*, 2021; Aston, 2015; Contreras and Christiansen, 2022; Thomson *et al.*, 2023). MWEs are combinations of words that often have a specific meaning that is different from the individual meanings of the words within the expression. These expressions are

treated as single units in language, and they can be idiomatic or have a fixed, non-compositional meaning. Here are some examples of multiword expressions:

- **Idioms:** These are expressions where the meaning is not directly related to the individual words. Examples include:
 - "A da din casă" (literally: to give from the house) - to reveal a secret.
 - "A bate câmpii" (literally: to beat the fields) - to talk nonsense.
- **Phrasal Verbs:** These are verbs combined with one or more particles (usually prepositions or adverbs) that create a unique meaning. Examples include:
 - "Give up" - meaning: to quit or surrender.
 - "Take off"- meaning: to remove, to fly or become successful.
- **Collocations:** These are combinations of words that tend to occur together frequently in a language. They often have a natural, customary pairing. Examples include:
 - "Strong coffee" (not "powerful coffee").
 - "Fast food" (not "quick food").
- **Proverbs:** These are short, traditional sayings that express a general truth or piece of advice. Here are some examples in Romanian:
 - "Cine seamănă vânt, culege furtună." - "Who sows the wind, reaps the whirlwind." (Similar to the English proverb "As you sow, so shall you reap.").
 - "Nu lăsa pe mâine ce poți face astăzi." - "Don't leave for tomorrow what you can do today." (Similar to the English proverb "Don't put off until tomorrow what you can do today.").
 - "Vorba dulce mult aduce." - "Sweet words bring much." (Similar to the English proverb "Speech is silver, but silence is golden.").
 - "Lupul își schimbă părul, dar năravul ba." - "The wolf may change its fur, but not its nature." (Similar to the English proverb "A leopard cannot change its spots.").

These proverbs reflect cultural wisdom and offer insights into human behavior and life experiences in the Romanian context.

- **Compound Words:** These are words formed by combining two or more words to create a new word with a specific meaning. Examples include:
 - "Nou-născut" (newborn baby).
 - "Software" (programs and data used on a computer).
- **Binomials:** These are pairs of words that are often used together and have a specific order. Here are some examples in Romanian:

- "Zi și noapte" - "Day and night".
- "Alb și negru" - "Black and white".
- **Set Phrases:** These are fixed expressions used in specific contexts. Examples include:
 - "A fost odată ca niciodată" - "Once upon a time" (used to begin a story).
 - "A da o mână de ajutor" - "To lend a helping hand".
- **Noun Phrases:** These are combinations of words that function as a single noun. Examples include:
 - "Carte de vizită" - "Business card".
 - "Masă rotundă" - "Roundtable".
- **Transitional Phrases:** These are expressions used to connect ideas or sentences in writing or speech. Examples include:
 - "În primul rând" - "First of all".
 - "Pe de altă parte" - "On the other hand".
 - "În plus" - "Furthermore", "In addition".
 - "Așadar" - "Therefore".
- **Technical Terms and Jargon:** In specialized fields, there are often multiword expressions that have specific meanings. For example:
 - "Protocol de comunicare" - "Communication protocol".
 - "Rețea de calculatoare" - "Computer network".

These are just a few examples of multiword expressions in language. They play a significant role in communication and can be quite diverse, ranging from everyday idioms to technical terms in various fields. In addition, MWEs exhibit several properties that distinguish them from regular word combinations. The properties of MWEs enhance their complexity in identifying them. Here are some key properties of multiword expressions:

- **Fixedness:** MWEs often have a fixed or rigid structure, and the words within the expression maintain a specific order. This fixedness contributes to the idiomatic nature of many MWEs.
- **Variability:** While some MWEs are highly fixed, others may exhibit a degree of variability, allowing for minor modifications in certain contexts.
- **Lexical Integrity:** MWEs often function as single lexical units, and altering any part of the expression may result in a change of meaning or loss of idiomaticity.
- **Semantic Unity:** The words within an MWE contribute to a unified, specific meaning. The combination of words is often chosen to convey a particular concept or idea.

- **Grammatical Unity:** MWEs exhibit grammatical cohesion, with the words in the expression functioning together as a syntactic unit. This cohesion contributes to the fixed syntactic structure of many MWEs.

These properties collectively contribute to the complexity and richness of multiword expressions in language. Researchers and language learners need to consider these characteristics to fully grasp the usage and nuances of MWEs in various linguistic contexts.

3 Augmented Reality in Language Learning

The research's innovation lies in applying Augmented Reality technology to language learning, especially in the context of MWEs. Augmented Reality can be a valuable tool for enhancing the learning and understanding of MWEs in various educational contexts. By integrating AR into language learning, educators and learners gain a powerful tool to bridge the gap between the abstract nature of MWEs and their practical use in real-world contexts. This research underscores the potential of AR technology to revolutionize language education, offering a compelling avenue for the mastery of Multiword Expressions and enhancing language proficiency in an increasingly interconnected world. In other words, language proficiency becomes crucial for effective communication in various international settings.

The AR tool developed operates in two primary modes: *“interactive visualization”* and *“real-world context integration”*. **Interactive Visualizations:** This mode harnesses Augmented Reality to offer dynamic visual representations of MWEs, aiding learners in comprehending the contextual and figurative meanings of these expressions through virtual imagery and animations. This innovative approach enhances the learning experience by providing engaging and immersive visualizations that go beyond traditional methods. Learners can explore the nuanced meanings of MWEs in a more interactive and memorable way, reinforcing their understanding through virtual depictions that bring linguistic concepts to life. Furthermore, as a member of Cost Action CA21167 Universality, diversity, and idiosyncrasy in language technology (UniDive, 2022), we are in the process of compiling a collection of Romanian Multiword Expressions (MWEs) that share similar meanings with their English counterparts. For example, the expression *“drink like a fish”* (see Figure 1) is an idiomatic phrase that is used to describe someone who consumes alcohol or drinks a lot of liquid, particularly alcohol, in large quantities. It implies excessive or heavy drinking, often to the point of intoxication.

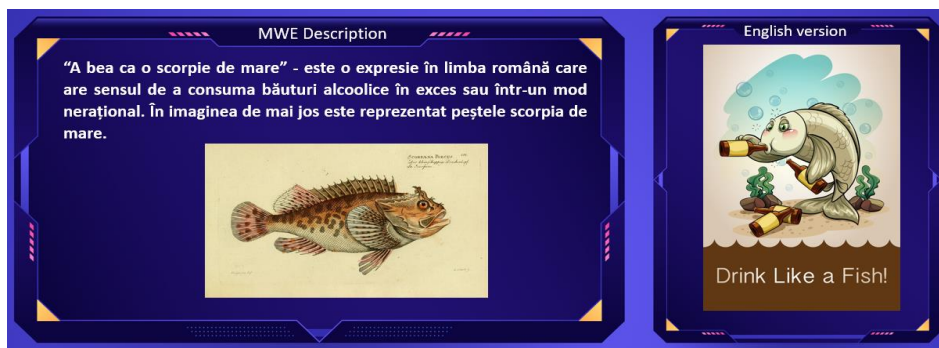


Figure 1. Romanian MWE “A bea ca o Scorpie de mare” illustration.

The origin of this expression likely comes from the behavior of some fish species, such as certain types of fish that continuously open and close their mouths, giving the impression that they are constantly "drinking" water. Therefore, saying that someone "drinks like a fish" humourously suggests that they consume liquids, especially alcoholic beverages, with a similar level of enthusiasm. It's important to note that this phrase is colloquial and may be used informally in conversation, but it is not a literal statement about someone's drinking habits. It's typically used in a lighthearted or humorous context to comment on someone's excessive or frequent alcohol consumption.

Real-world Context Integration: In this application, AR superimposes Multiword Expressions onto physical markers within real-world scenarios. This approach provides an immersive learning experience, enhancing understanding and retention by connecting expressions to relevant contexts. Learners can interact with MWEs in their everyday environment, reinforcing the association between these expressions and the practical situations in which they are used. This integration of AR and real-world contexts not only facilitates a more engaging learning experience but also deepens the comprehension of MWEs by grounding them in authentic and meaningful settings (see Figure 2).

4 Bridging the Gap

Integrating AR in language learning effectively bridges the theoretical and practical aspects of MWEs in everyday communication. By providing a contextual and interactive learning experience, AR simplifies the complexities associated with these expressions, making them more accessible to learners. Augmented Reality can be a valuable tool for enhancing the learning and understanding of Multiword Expressions in various educational contexts. Here are some ways in which AR can be used to facilitate the learning of MWEs.



Figure 2. List of Romanian MWEs related to “time is running”.

- **Gamification:** Gamifying the learning process through AR can make it more engaging and fun. Gamified AR apps can challenge learners to identify and use MWEs in real-world scenarios, increasing motivation and retention.
- **Language translation and transliteration:** AR can assist learners in translating MWEs into their native language or provide phonetic pronunciation guides, enhancing their understanding and pronunciation of these expressions.
- **Flashcards and quizzes:** AR-based flashcards and quizzes can be used to reinforce the learning of MWEs. Learners can scan MWE flashcards or objects, and the AR app can provide definitions and usage examples.
- **Real-time feedback:** AR apps can provide real-time feedback on the correct usage of MWEs, helping learners refine their language skills and gain confidence in using these expressions.
- **Enhanced Pronunciation Practice:** AR applications can provide instant feedback on pronunciation by analyzing learners' speech in real time. This feature helps refine pronunciation skills, offering a personalized and effective way to work on language sounds and intonation.
- **Adaptive Learning Environments:** AR adapts to individual learning styles and paces, providing a personalized learning experience. Learners can receive tailored exercises, feedback, and content based on their progress and areas that need improvement.
- **Collaborative Language Learning:** AR facilitates collaborative language learning experiences, allowing learners to interact with virtual elements

together. This collaborative aspect fosters communication and teamwork, essential components of language acquisition.

By integrating AR into language learning, teachers can create immersive and engaging experiences that help learners better understand, remember, and use Multiword Expressions effectively in real-world situations. AR's ability to provide context, interactivity, and personalization can significantly enhance the learning process for MWEs and other aspects of language acquisition.

5 Implications and Future Direction

The use of AR in language education introduces new, engaging, efficient, and effective learning methodologies. It presents a fresh approach to mastering MWEs, essential for fluency and natural language use. The study highlights AR's potential to revolutionize language education, particularly significant in a globally interconnected environment where linguistic proficiency is key (Parmaxi and Demetriou, 2020; Cai *et al.*, 2022; Chua and Yu, 2023). Our upcoming initiatives involve collaboration with Dr. Elva Leca from the Polytechnic University of Tirana, Albania, and Dr. Lucía Amorós from the University of Murcia, Spain. Together, we aim to develop a multilingual application focused on learning Multiword Expressions (MWEs). This collaborative effort seeks to enhance language learning experiences by incorporating a diverse range of expressions from different linguistic backgrounds. The application will provide a valuable resource for learners looking to expand their understanding of MWEs in various languages, fostering cross-cultural exchange and linguistic proficiency.

6 Conclusion

In conclusion, our research represents an exploration at the intersection of technology and language learning, specifically focusing on the use of Augmented Reality (AR) to enhance the comprehension and application of Multiword Expressions (MWEs) in the Romanian language. The study, led by researchers from the Vladimir Andrunachievici Institute of Mathematics and Computer Science and the Technical University of Moldova, underscores the transformative potential of AR technology in language education. The research delves into the intricacies of Multiword Expressions, emphasizing their importance in everyday language use. Recognizing the challenges these expressions pose in language acquisition due to their non-literal and idiosyncratic nature, our team compiled an extensive corpus of 17,200 Romanian MWEs. This research is a contribution to the fields of language learning and technology. By utilizing Augmented Reality, we offer a novel and compelling approach to understanding and mastering

Multiword Expressions. This study underscores the importance of MWEs in language proficiency and the transformative potential of AR in education, setting a precedent for future advancements in digital learning methodologies.

References

- Espinal, M., and Mateu, J. 2019. *Idioms and Phraseology*. Oxford Research Encyclopedia of Linguistics. Retrieved 22 Nov. 2023, from <https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-51>.
- Vercellotti, M., Juffs, A., and Naismith, B. 2021. Multiword Sequences in English Language Learners' Speech. *System*, 2021, 98. 102494. 10.1016/j.system.2021.102494.
- Aston, G. 2015. Learning Phraseology from Speech Corpora. *Studies in Corpus Linguistics*, 69, 63-84. <https://doi.org/10.1075/scl.69.04ast>
- Contreras K. P. and Christiansen M.H. 2022. Models of Language and Multiword Expressions. *Front Artif Intell*. 2022 Feb 17; 5:781962. DOI: <https://doi.org/10.3389/frai.2022.781962>. PMID: 35252848; PMCID: PMC8892141.
- Thomson, H., Coxhead, A., Boers, F., and Warren, P. 2023. Increasing Use of Multi-word Expressions in Conversation through a Fluency Workshop. *Systems* Vol.113, April 2023, 102294, 10.1016/j.system.2023.102994
- UniDive. 2022. Cost Action CA21167. *Universality, Diversity and Idiosyncrasy in Language Technology* <https://www.cost.eu/actions/CA21167/#tabs+Name:Description>
- Parmaxi, A. and Demetriou, A. A. 2020. Augmented Reality in Language Learning: a State-of-the-Art Review of 2014–2019. *Journal of Computer Assisted Learning*, 36(6), 861-875. <https://doi.org/10.1111/jcal.12486>
- Cai, Y., Pan, Z., and Liu, M. 2022. Augmented Reality Technology in Language Learning: a Meta-Analysis. *Journal of Computer Assisted Learning*, 38(4), 929-945. <https://doi.org/10.1111/jcal.12661>
- Chua, H.W. and Yu, Z. 2023. A Systematic Literature Review of the Acceptability of the Use of Metaverse in Education over 16 Years. *J. of Computers in Education*, <https://doi.org/10.1007/s40692-023-00273-z>
- Ilinca, V. 2015. *Dicționar de expresii românești în context (DERC) A-C*. Presa Universitară Clujeană, p.v304. ISBN 978-973-595-856-5.

CHAPTER 3

ANONYMISATION AND SUMMARISATION

TOWARDS ANONYMISING THE ROMANIAN JURISPRUDENCE

VASILE PĂIȘ¹, VERGINICA BARBU MITITELU¹, ELENA IRIMIA¹,
RADU ION¹, VALENTIN BADEA¹,
DAN TUFIȘ¹, COSMIN STEREA-GROSSU²

¹Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Bucharest, Romania

{vasile, vergi, elena, radu, valentin.badea, tufis}@racai.ro

²Superior Council of Magistracy, Bucharest, Romania

cosmin.sterea@csm1909.ro

Abstract

The transparency of the judicial process and the consistency of judicial decisions can be improved through their publication, which requires the anonymisation of documents. This can be done by applying specific technical solutions that can help to automate the preparation of the document for publication. More than 40 million decisions have been published in the Romanian case law database, available in the ReJust portal, developed by the Superior Council of Magistracy. The initial anonymisation system employed only basic methods, such as name lists and regular expressions. This paper discusses new approaches that will be applied to anonymise Romanian jurisprudence using state-of-the-art technologies, including large language models and deep neural networks.¹

1 Introduction

A classical definition of jurisprudence was formulated during the reign of the Roman Emperor Justinian as “Juris prudentia est divinarum atque humanarum rerum notitia, iusti atque iniusti scientia” (Iustiniani Institutiones, 2002) (The jurisprudence is the knowledge of divine and human things, the science of what is right and wrong). In other words, this doctrine of Roman law presumed not only the knowledge of the “Divine law” and Human law, but also the possession of the science of what is right and unjust, that is, of justice and injustice (Dură, 2019). Tur (1978) acknowledges that in the English language, the term “jurisprudence”

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

is used interchangeably for “philosophy of law” and “legal theory”, but it sometimes appears in a sense equivalent to the French term “la jurisprudence”, meaning case-law, or that part of the positive or actual law built up by judicial decision rather than enacted and laid down in statutes or in codes: what Bentham called “judge-made law” (Tur, 1978; Postema, 2019).

Access to a Romanian case-law database will benefit public prosecutors, lawyers, notaries, law professionals, law professors, and students for legal and educational research purposes. Judges and legal advisors from lower courts in Romania will consult information and documents published in the jurisprudence database to ensure that their judgements are harmonised with the case law of the higher instance courts. Furthermore, instant access to relevant case law fosters harmonisation of court practice. In this context, the Superior Council of Magistracy (CSM) developed the jurisprudence portal ReJust² to allow citizens and practitioners within the judicial system easier access to court decisions issued by national courts. Access to the jurisprudence portal is free with mandatory user registration. The case law database contains all court decisions from criminal, civil, commercial, and administrative case types. The only exceptions from publication are documents marked in ECRIS³ (the case management system) as confidential, such as the judgements or other decisions related to minors, sexual harassment, divorces/family matters, decisions acknowledging mediation settlements, verdicts on offences as treason, espionage, rape and child pornography and upon a request of a party in the proceedings. Case law database contains judgements from all courts except the High Court of Cassation and Justice, which maintains its own case law database. Currently, approximately 40 million court decisions issued from 2011 to the present are published in this portal, and the database is updated daily.

Opening access to court decisions to a larger audience means anonymisation has to be employed to protect the parties involved. Data anonymisation is the process of removing identifiable personal information from collected data. Article 4, paragraph 5, of the General Data Protection Regulation (GDPR)⁴, defines “pseudonymisation” as the processing of personal data in such a manner that it can no longer be attributed to a specific data subject without the use of additional information. Other legal documents define anonymisation requirements in general and in the specific case of the case law database, such as Law no. 190/2018 on implementing measures to Regulation (EU) 2016/679 of the European Parliament

² <https://www.rejust.ro/>

³ https://commission.europa.eu/law/cross-border-cases/judicial-cooperation/tools-judicial-cooperation/european-criminal-records-information-system-ecris_en

⁴ Council Regulation 2016/679, 2016 O.J. (L119) (EU), <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data and repealing Directive 95/46/EC (General Data Protection Regulation), and Decision of the Section for Judges of the Superior Council of the Magistracy no. 998/17.03.2022.

This paper introduces the project “System for the anonymisation of Romanian jurisprudence (SARoj)”⁵. It describes work-in-progress related to implementing the anonymisation solution that will become part of the ReJust portal. The rest of this paper is organised as follows: Section 2 presents related work, Section 3 describes the corpus creation and annotation, Section 4 introduces the envisaged system architecture, and Section 5 presents conclusions and future work.

2 Related Work

The success of the AI systems based on neural networks and machine learning generated a global rush for collecting more and diverse data, as clean as possible and compulsorily GDPR compliant. Anonymisation, as mentioned before, is the process of removing confidential information (named entities, NE), such as person or organisation names, locations, dates, personal code identifiers (CNP), phone numbers, car plates, emails, account numbers, etc. While expert manual anonymisation, supposedly, is a robust and trustful option in case of a small number of documents, for large volumes of texts (e.g. millions of documents) it becomes impracticable. Therefore, the process should be automated with human intervention minimised. It means the process should be able to automatically recognize the named entities (NER) that represent confidential information and afterwards anonymise them according to several available techniques. These two steps, together with their prerequisites, will be discussed in the next sections.

The anonymisation process became one of the top priorities for public administration, research and industry. Machine translation is one of the major areas that need vast amounts of data, and, when collected from the internet, it requires anonymisation before being put into production. The European Commission, Digital Service Infrastructures of the European Commission (EC), as well as owners and creators of huge amounts of data that contain sensitive information, must anonymise the documents before releasing them to the public. The comprehensive report from CrossLang (Vanallemeersch and Szoc, 2021) reviews some of the European projects that delved into the anonymisation process and discusses available (public and commercial) tools for anonymisation. Some of the European actions which addressed anonymisation are: ELRC⁶, CEF Data

⁵ <https://www.racai.ro/p/saroj/>

⁶ <https://www.lr-coordination.eu/>

Marketplace⁷, COMPRISE⁸, ParaCrawl⁹, MAPA¹⁰, MARCELL¹¹ (Váradi *et al.*, 2020; Tufiș *et al.*, 2020), CURLICAT¹² (Váradi *et al.*, 2022), etc. Besides the tools created by the funded projects of these actions (e.g. TM-Anonymiser, TEXT-Transformer, Biroamer, LegalNERo), some other institutions (National Library of Medicine, National Institutes of Health of the United States), companies (SDL, SYSTRAN, ABMartin, STACC, TILDE, and others), research organisations (CNRS, Barcelona Supercomputing Center, DFKI, most academies and universities in CEE Europe) have been involved in projects on building/collecting language resources, requiring anonymisation before dissemination. The anonymisation was performed either by using available tools or by building language-specific tools. Some of the current monolingual or multilingual anonymisation software are free, others are commercial products: NETANOS¹³, NLM-Scrubber¹⁴, MEDINA¹⁵, SDL TM Anonymizer¹⁶, SYSTRAN Anonymizer for Relativity¹⁷.

Most of the anonymisation tools deal with general language or were meant for specific domains (e.g. health). There are few available tools for the legal domain and they are usually language specific.

The SAROJ project presented in this paper has the objective of building an anonymisation instrument for Romanian jurisprudence. Stimulated by European institutions (EC, EU, Council of Europe, DG JUST, etc.), similar projects were launched in almost all European countries (see the round table and the webinars on the use of Artificial Intelligence in the justice field¹⁸). One of the most successful projects funded by EC is MAPA (Multilingual Anonymisation for Public Administrations - an INEA-funded Action for the European Commission under the Connecting Europe Facility). It covers all EU official languages and is open-source¹⁹. MAPA deals with public administration languages. The SAROJ project is considering a more specialised language (Romanian jurisprudence) and

⁷ <https://www.datamarketplace.eu/>

⁸ <https://www.compriseh2020.eu/>

⁹ <https://paracrawl.eu/>

¹⁰ <https://mapa-project.eu/>

¹¹ <https://marcell-project.eu/>

¹² <https://curlicat-project.eu/>

¹³ <https://github.com/ben-aaron188/netanos>

¹⁴ <https://lhncbc.nlm.nih.gov/scrubber/announcement.html>

¹⁵ <https://medina.limsi.fr/>

¹⁶ <https://appstore.rws.com/Plugin/98>

¹⁷ <https://www.systransoft.com/download/brochures/systran-relativity-anonymizer.pdf>

¹⁸ https://commission.europa.eu/strategy-and-policy/policies/justice-and-fundamental-rights/digitalisation-justice/conferences-and-events_en

¹⁹ https://gitlab.com/MAPA-EU-Project/mapa_project

a huge number of documents representing court decisions issued by judges, tribunals, and courts of appeal.

The problem is critical for institutions that require or plan to publish their data. The major IT companies and research community are also concerned with the anonymisation process for their collected huge amounts of data necessary for building large language models and various applications.

3 Corpus Creation

The large language models and the deep neural networks for anonymisation developed in our project are trained and tuned on a manually annotated corpus. The corpus comprises files that represent the official statements of the punishment given by a law court to a guilty person or organisation. Their structure is predefined: the file header identifies the case number, contains the name of the country, the county, and the court of justice where the trial took place, alongside the date when the statement was issued, the names of the court president and clerk; the body of the file is a narrative which presents the parts in the case, the names of their lawyers, of the prosecutors, as well as the presentation of the facts under analysis, the laws applicable to the case and the extent to which they were obeyed or disobeyed by the parties; the last part of the narrative contains the statement of the verdict, and the date and the names of the president and clerk are repeated.

The manually annotated files are selected from various courts throughout the country to ensure the diversity of the cases and, implicitly, of the NEs they contain. The annotation process strictly follows specially created guidelines, in which 16 types of NEs are defined and described:

- PER (abbreviation for PERSON) entities include human names, surnames, and nicknames. A person can be an individual or a group (e.g. “familia [**Popescu**]²⁰” / En. “f[**Popescu**] family“f). Annotations²¹ take in consideration the largest sequence of text that clearly identifies the name, without additional markers like titles, functions (e.g: “Preşedintele Agenţiei Naţionale de Administrare Fiscală [**Mioara Oprea**]“/En. ”President of the National Fiscal Administration Agency [**Mioara Oprea**]”) or group markers (see “family” in the example before). All person entities in the documents are annotated, regardless of their role in the trial. When more names are associated with the same person (like “[**Stan Ion zis Viperă**]” / En. “[**Stan Ion called the Viper**]” or “[**Mitroi**

²⁰ All named entities in examples are fictitious, for anonymisation purposes.

²¹ The NE that is annotated is boldfaced and placed between square brackets in the examples we give here.

Ana, fostă Petre]”/ En. **”[Mitroi Ana, formerly Petre]”**) they are annotated as a single entity if they are adjacent to each other. In a situation like **”[Radu Popescu], fiul lui [Matei] și al [Ioanei]”** / En. **”[Radu Popescu], son of [Matei] and [Ioana]”**, there are three different entities in the sequence.

- INITIALS entities refer to mentions of the initials of the court president and clerk signing the statement, located at the end of the document. Examples: **”Red. [RMI]/Tehnored. [VP]/2ex”**, **”Tehnored. Jud. [P.A.T.]”**.
- ORG (abbreviation for ORGANISATION) entities include companies, corporations, agencies and any other group defined by a formal organisational structure. The annotation process will mark the text sequences that clearly indicate the name of an organisation, without other adjacent words or entity markers (like SC, S.R.L, etc.): e.g., **”compania SC Company SA”**, **”[Compania Electricitate Brașov]”**. In the first example, the word **”compania/ En. company”** is not part of the organisation's name and is not annotated. In contrast, in the second case, the same word is part of the organisation's name and is annotated. In any situation where the annotator is in doubt about which sequence to annotate, the full name of the organisation can be Googled. All private organisations are annotated, whatever their quality in the trial, while public organisations are annotated only if they are parts in the trial.
- LOC (abbreviation for LOCATION) represents, in our project, all geographic locations that can identify people or organisations involved in the trial: domiciles, headquarters, addresses of properties in dispute, descriptions of terrain locations and neighbourhoods, birthplaces, including locations from abroad. Addresses where accidents or events take place are not annotated. Other location elements that could, by association with other non-anonymised information in the document or information available online, lead to the identification of the individuals or companies involved must be annotated: e.g. **”Proprietarul restaurantului [Orhideea]/En. The owner of the restaurant [Orhideea]”**, **”Reclamantul solicită retrocedarea unei suprafețe de 1000 ha din [Parcul IBB]. / En. The plaintiff requests the retrocession of an area of 1000 ha from the [IBB Park]”**. Annotations for LOC entities include the entity markers, like in: **”cu domiciliul în [localitatea Oltenița, Str. Gării Nr. 1 Bl. A1 Ap 1 Et. 1]/ En. with domicile in [Oltenița, Str. Gării Nr. 1 Bl. A1 Ap 1 Et. 1]”**.
- DATE is used to annotate time expressions that fully identify a person's date of birth: **”Născut în [19.01.1988]/ En. Born on [19.01.1988]”**.

- DECISION, CASE and DOCNUMBER refer to identification numbers specific to documents: DECISION for court decisions, CASE for court files and DOCNUMBER for any other document mentioned in the files that could lead to the identification of the parties (reports, mandates, receipts, contracts, decisions of other institutions, etc.). General decisions and judgments (issued by the CCR, CJEU, ECHR etc.) that do not refer specifically to the parties, will not be annotated. Only the actual number of the document (the string in front of the first "/" separator) is annotated: "Hotărârea nr. **[123]/03.08.2022** / En. Decision no. **[123]/03.08.2022**", "[**3399**]/105/2023", "Dosar nr.[**5**]/J/2014 / En. File no. [**5**]/J/2014".
- CNP refers to the unique personal numerical code, consisting of 13 digits, assigned to each person when their birth is registered in the civil status registers (e.g. "CNP **[1750214XXXXX]**").
- ID is used to annotate personal identification document numbers and series. ID covers documents such as identity cards, provisional identity cards, passports, travel documents, residence permits, birth certificates, driving licences, etc., i.e. all personal documents that can identify someone's identity. If the document has both series and number, the annotated sequence will start after the word "seria": ex. "Ion Marin, legitimat cu CI seria **[AG nr. 123456 eliberat la data de 15.11.2012 de SPCEP jud Argeş]**/ En. "Ion Marin, registered with CI series **[AG no. 123456 issued on 15.11.2012 by SPCEP Argeş county]**."; otherwise, it will start after the word "nr.": e.g. "Ion Marin, legitimat cu paşaport diplomatic nr. **[123456789 eliberat la data de 10.12.2010]**/ En. "Ion Marin, registered with diplomatic passport no. **[123456789 issued on 10.12.2010]**."
- EMAIL and PHONE are used to annotate the email addresses and phone numbers that could disclose the identity of the trial parties. Entity markers like "email", "tel.", will not be part of the annotations.
- ECLI is used for European Case Law Identifiers and has the same easily recognizable format for all EU Member States and courts. In Romania, the ECLI code consists of the following elements, separated by a colon: "ECLI"; country code "RO"; court code; the year of the decision; a sequence number which consists of the internal number of the corresponding electronic register in the ECRIS system (unique at the level of the court) and the number of the judgement, as it is given in the register of judgments of the court (unique at the level of the register). The annotated sequence starts with the court code. Example: ECLI:RO:[**JDS5B:2017:012.012345**].

- CUI stands for Commercial Register Identification numbers (abbreviated CUI in Romanian and used before 2007), for Fiscal Identification Code (replacing CUI after 2007) and for order numbers in the Commercial Register. These are identifiers associated with businesses, cooperatives, companies, etc. registered in the trade register. Entity markers are not included in the annotation: “CIF [**RO123435678C**]”, “Trade register registration number [**J12/34/2020**]”.
- IBAN (International Bank Account Number) is used to annotate sequences of up to 34 letters and numbers that identify the country, verification number, bank location and bank details of a specific bank account. Only the actual numbers/accounts are annotated, without the “IBAN” marker word: IBAN [**RO49 AAAA 1B31 1234 5678 0000**].
- NCAD is used for numbers from the real estate register: cadastral/topographic numbers, registration numbers in the land register, identification codes in the cadastre (which include the SIRUTA²² code of the county, followed by the intra/extravilla code (1=extra, 2,3,4 etc.=intra), followed by cadastral number). The actual numbers/codes are annotated in their entirety, but without accompanying marker words such as “no.”, “topographical number”, etc.: ”Înregistrat în CF nr. [**12345-C1-Z12**] a localității Crevedia, cu număr topografic [**123/4/5**]”/En. ”Registered in CF no. [**12345-C1-Z12**] of the town of Crevedia, with topographic number [**123/4/5**]”.
- AUTO is used for car plates and chassis series: ”automobil cu număr de înmatriculare [**B 12 ABC**] și serie de șasiu [**SV12-1234567**]”/ En. ”car with registration number [**B 12 ABC**] and chassis series [**SV12-1234567**]”.

38 annotators (experts from both the linguistic and judicial fields) are currently involved in the annotation process. In the current phase, 1,500 documents are considered for annotation, with some of them having double annotation, so that an inter-annotator agreement score can be calculated to show the annotation reliability.

The annotation was done on a dedicated instance of the RELATE platform (Păiș *et al.*, 2020; Păiș, 2020; Păiș *et al.*, 2019), which allows the user to manually select the string to be annotated by opening a pop-up window displaying the list of predefined NEs for the user to choose from (see Figure 1). Given that some NEs occur several times in the same file, the platform offers the possibility of

²² SIRUTA = Sistemul Informatic al Registrului Unităților Teritorial - Administrative (The IT System of the Register of Territorial - Administrative Units in Romania)

automatically annotating all occurrences of the same NE when first encountered in the file, which is quite timesaving.

The RELATE platform integrates the BRAT Rapid Annotation Tool²³ (Stenetorp *et al.*, 2012) as a control allowing the highlighting of text spans representing the entities. The annotators use the mouse to select a text span and a popup allows associating of an entity type. In case of errors, it is possible to edit the selected span or delete it. Entity definition and storage of resulting annotation files is handled within the RELATE platform. Once saved, a file can be retrieved at any time for further annotation or corrections, if necessary.

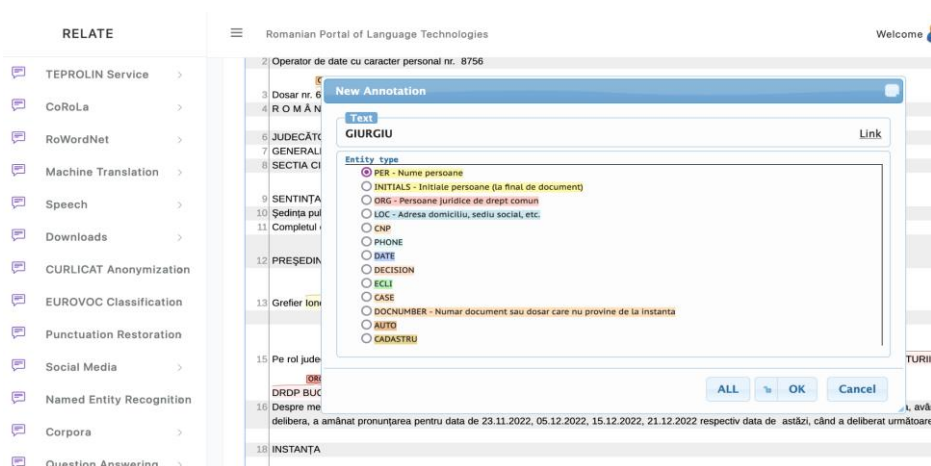


Figure 1. The BRAT annotation tool integrated in the RELATE platform, used for manual annotation of the named entity corpus.

4 Envisaged System Architecture

The system architecture is designed to be a versatile and powerful solution for processing and manipulating textual data. At its core, it comprises at least eight interconnected module types, each playing a unique role within the data processing pipeline. These modules facilitate token-based anonymisation and document reconstruction in the original format. The modules are designed as REST API endpoints, communicating over the HTTP protocol, with parameters passed in JSON format. Modules employing token-based annotations work with a column-based CoNLL-U Plus²⁴ file format. Annotations are always added on

²³ <https://brat.nlplab.org/index.html>

²⁴ <https://universaldependencies.org/ext-format.html>

the last column of the file. Modules providing named entity recognition (NER) employ a BIO (Begin/Inside/Outside) annotation format.

Multiple instances of the same module may be started with different parameters (different dictionaries or employed algorithms) in order to construct a complex annotation pipeline. An orchestration service is being developed to allow automatic start of different modules and pipeline-based processing. A new document is given to the orchestration service, which in turn passes it along each pipeline step and collects results. In case of errors, the orchestration service stops the processing and returns a predefined error message, together with additional information from the module that caused the error.

Each implemented module will provide a dedicated API for reporting its current status. This information can be globally interrogated from the orchestration service using a single API, called “checkHealth”. This API calls each individual module’s endpoint and aggregates the result, providing it back to the caller. This will provide early warning in case of malfunctions and information on the available resources (such as number of dictionary entries, number of predefined entities).

In its initial stage, the following modules are considered:

- (1) **Text Extractor** module to transform documents from their original format (DOCX) into the structured CoNLL-U Plus. This format includes tokenized text, basic language annotations (lemma, part-of-speech, dependency parse tree) and the positions of each token in the original document, allowing for easy reconstruction of the original content once the replacement token is identified. This module is essential for converting unstructured text into a structured format suitable for further analysis.
- (2) **Regex Annotator** adds named entity tags to the tokens based on a list of regular expressions. These are particularly useful for well-known structured entities, such as CNP, IBAN.
- (3) **Dictionary** enhances CoNLL-U Plus files with entity annotations using a predefined dictionary of known entity names of values.
- (4) **Model-based named entity recognition** performs NER using a Transformers-based Large Language Model (LLM) with the standard NER fine-tuning procedure (Devlin *et al.*, 2019). The LLM employed in current experiments is a Romanian BERT model that is being developed specifically for Romanian and trained on the Corpus of Contemporary Romanian Language (CoRoLa) (Barbu Mititelu *et al.*, 2019). This implies the development of a custom Romanian WordPiece tokenizer that performs text normalisation (diacritic and whitespace replacement and automatic enforcement of the Romanian Academy writing norms) and proper Romanian clitic splitting, beside splitting at whitespace. The tokenizer’s

vocabulary will be much larger than the typical BERT English vocabularies of around 30,000 tokens, to accommodate an inflected language such as Romanian.

- (5) **Voting** processes multiple CoNLL-U Plus files (generated by different modules or by the same module with different parameters, such as different dictionaries), each containing identified entities in the same source document, and applies a voting algorithm. Currently 4 algorithms are considered: DIFF - computes a difference between files; ADD - combines multiple files by keeping any annotation; INTERSECT - computes the intersection between the annotations; MAJORITY - provides majority voting between multiple files.
- (6) **Entity Encoding** provides encoding of entities recognized in these documents using unique identifiers, including Romanian language-specific suffixes (i.e., Maria => #PER1, Mariei => #PER2_ei, Clujului => #LOC3_ului, etc.). This encoding mechanism follows the previous work of Păiș *et al.* (2021b). Preservation of suffixes ensures that in a later step, entities could be replaced with similar-looking words, in the case of a pseudonymisation process. Entities present in documents belonging to the same case will be encoded using the same identifiers.
- (7) **Entity Mapping** performs replacement of encoded entities following a predefined scheme. This includes dictionary-based pseudonymization or anonymisation based on special characters (such as “#”). In the first case, a dictionary of possible replacement candidates is provided with entries for each entity type. If an entity type is not found in the dictionary, the algorithm automatically falls back to the second option, replacing the entity with a random number of customizable characters (i.e., X, #).
- (8) **Document Reconstruction** handles the final step in the pipeline. This module generates anonymised versions of original documents. It takes the final CoNLL-U Plus file and the original document as input. It produces a new document in which the original text is replaced with anonymised content, facilitating the secure sharing of sensitive data. The original document formatting (such as bold, italic, paragraphs) is preserved.

At the heart of these modules are the CoNLL-U Plus files that encompass all the information needed for text processing. The first 10 columns are those in a regular CoNLL-U file (ID, Form, Lemma, UPos, XPos, Feats, Head, Deprel, Deps, Misc). Additional columns are employed for: START- start position in the original document, END - end position in the original document, NER - the BIO token-based named entity annotations, NER_ID - the entity identifier produced by the Entity Encoding module, ANONYMIZED - the anonymised string associated with the entity. An example is given in Figure 2.

1	Domiciliul 5	domicili nsubj	NOUN	Ncmsry	Case=Acc,Nom Definite=Def Gender=Masc Number=Sing 1439 1449	O	—	—
2	actual 1	actual amod	ADJ	Afpms-n	Definite=Ind Degree=Pos Gender=Masc Number=Sing 1450 1456	O	—	—
3	al 4	al det	DET	Tsms	Gender=Masc Number=Sing Poss=Yes PronType=Prs 1457 1459	O	—	—
4	Oanei 1460	Oanei 1465	PROPN B-PER	Np #PER1_ei	— Mariei	1	nmod	—
5	este 1466	fi 1470	VERB	Vmip3s Mood=Ind Number=Sing Person=3 Tense=Pres VerbForm=Fin 1466 1470	O	—	—	—
6	în —	în —	ADP	Spsa 1471	AdpType=Prep Case=Acc O	—	—	7 case
7	Cluj SpaceAfter=No	Cluj SpaceAfter=No	PROPN 1475	Np 1479	— B-LOC	5 #LOC2	obl #####	—
8	.SpacesAfter=\s\n	.SpacesAfter=\s\n	PUNCT	PERIOD 1479	— O	5 —	punct	—

Figure 2. Example CoNLL-U Plus file.

Basic language processing, within the Text Extractor module, is currently 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). However, depending on the final processing time, other implementations will be considered, such as NLP-Cube (Boroș *et al.*, 2018) or Stanza (Qi *et al.*, 2020), which may provide increased accuracy for the basic annotations (yet unknown how this will reflect in the final NER annotations) with more processing time required. Models for different basic language resource kits, trained on the same RRT 2.7, are already available within the RELATE platform²⁵.

The system is designed to work in two scenarios: (a) complete anonymisation of the entire database (or large batches); (b) anonymisation of a single document. The batch implementation is required for the initial anonymisation process and for any additional re-anonymisation when the system models or resources are updated. Furthermore, it will also be used for anonymisation of new batches imported in the database. However, in special cases single documents may need to be quickly anonymised. For example, in the case of a system update requiring re-anonymisation of the entire database, a user may try to access a document before it was re-anonymised by the batch processing mechanism. In this case, the single document pipeline will be used to provide quick anonymisation before sending the document to the user. These features are implemented in the orchestration web service and are independent of the individual module's implementation.

²⁵ https://relate.racai.ro/index.php?path=repository/resource&resource=annotation_models_rrt_2_7

5 Conclusion and Future Work

This paper introduced the anonymisation system for Romanian jurisprudence, which is still under development. Its modular architecture (see Section 4) is designed to be easily extensible. Each module implements a specific function and can be upgraded independently or replaced with a different implementation. When finalised, it will allow anonymising the content of the ReJust²⁶ portal, offering access to the Romanian jurisprudence. Development for the SAROJ project happens in the GitHub²⁷ of the Institute for Artificial Intelligence of the Romanian Academy. All publicly available developments are included in the repository. The project's components are designed to be re-usable, with the complete processing pipeline offering a single, integrated, web service. It is designed to be available in a containerized environment, by creating a Docker-based container.

Future work focuses on finishing the implementation of the different modules, detailed in Section 4, as well as completion and validation of the corpus, detailed in Section 3, used for training artificial intelligence models. When completed, the entire pipeline needs to be evaluated against a “test” subset of the entire corpus. Finally, different parameters will require fine-tuning to improve the system's performance. The end-system, even though constructed for anonymising jurisprudence, will be adaptable to other anonymisation tasks as well, given proper resources and neural models.

Acknowledgements

This work was performed within the project “System for the anonymisation of Romanian jurisprudence (SAROJ). The SAROJ project was carried out with funding from the Council of Europe, within the project "Foster Transparency of Judicial Decisions and Enhancing the National Implementation of the ECHR (TJENI)" funded by the EEA and Norway Grants Fund for Regional Cooperation. The views expressed herein can in no way be taken to reflect the official opinion of the Council of Europe.

References

- Iustiniani Institutiones*. 2002. Text latin și traducere în limba română, cu note și studiu introductiv de prof. dr. doc. Vladimir Hanga. Lumina Lex. București, Romania.
- Barbu Mititelu, V. 2018. Modern Syntactic Analysis of Romanian. In *Clasic și modern în cercetarea filologică românească actuală*, Ichim, O. (coord.), Botoșineanu, L.,

²⁶ <https://www.rejust.ro/>

²⁷ <https://github.com/racai-ai/saroj>

- Butnaru, D., Clim, M.R., Ichim, O., Olariu, V. (eds.), Editura Universității „Alexandru Ioan Cuza”, Iași, pages 67-78.
- Tușiș, D., Barbu Mititelu, V., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., and Onofrei, M. 2019. Little Strokes Fell Great Oaks. Creating Corola, the Reference Corpus of Contemporary Romanian. In *Revue roumaine de linguistique*, No./Issue 3.
- Boroș, T., Dumitrescu, Ș., and Burtica, R. 2018. NLP-Cube: End-to-End Raw Text Processing with Neural Networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics. pages 171-179.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186, <http://dx.doi.org/10.18653/v1/N19-1423>.
- Dură, N. V. 2019. „Jus” (Law) and „Justitia” (Justice) in Roman Solicitors’ Perception and Definition. Reflections and Evaluations. *Logos Universality Mentality Education Novelty: Law*, 7(2), 45-56. <https://doi.org/10.18662/lumenlaw/26>.
- Postema, G. 2019. *Bentham and the Common Law Tradition*, 2nd ed, Clarendon Law Series, Oxford, 2019, online ed, Oxford Academic, 19 Sept. 2019. <https://doi.org/10.1093/oso/9780198793052.001.0001>.
- Păiș, V., Tușiș, D., and Ion, R. 2019. Integration of Romanian NLP Tools into the RELATE Platform. In *Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing, ConsILR 2019*. pages 181-192.
- Păiș, V., Ion, R., and Tușiș, D. 2020. A Processing Platform Relating Data and Tools for Romanian Language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, Marseille, France, European Language Resources Association, pages 81-88.
- Păiș, V. 2020. Multiple Annotation Pipelines Inside the RELATE Platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing, ConsILR 2020*. pages 65-75.
- Păiș, V., Ion, R., Avram, A.-M., Mitrofan, M., and Tușiș, D. 2021a. In-depth Evaluation of Romanian Natural Language Processing Pipelines. In *Romanian Journal of Information Science and Technology (ROMJIST)*, vol. 24, no. 4, pages 384-401.
- Păiș, V., Irimia, E., Ion, R., Tușiș, D., Mitrofan, M., Barbu Mititelu, V., Avram, A.-M., and Curea, E. 2021b. Romanian Text Anonymisation Experiments from the CURLICAT Project. In *The 16th International Conference on Linguistic Resources and Tools for Natural Language Processing, ConsILR 2021*. pages 165-178.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Association for Computational Linguistics (ACL) System Demonstrations, 2020*.
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. 2012. Brat: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL*.

- Straka, M., Hajič, J., Straková, J. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.
- Tufiș, D., Mitrofan, M., Păiș, V., Ion, R., and Coman, A. 2020. Collection and Annotation of the Romanian Legal Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 2766-2770.
- Tur, R. H. S. 1978. *What is Jurisprudence? The Philosophical Quarterly*, 28(111), 149-161, Oxford University Press, <https://doi.org/10.2307/2219360>.
- Vanallemeersch, T. and Szoc, S. 2021. Final Report D91.1 Specification on Anonymisation. *SMART 2019/1083 Action on CEF Automated Translation Core Service Platform*, 126 pages.
- Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pezik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., and Brank, J. 2020. The MARCELL Legislative Corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 3754-3761.
- Váradi, T., Nyéki, B., Koeva, S., Tadić, M., Štefanec, V., Ogrodniczuk, M., Nitoń, B., Pezik, P., Barbu Mititelu, V., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., and Repar, A. 2022. Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources. In *Proceedings of the Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pages 100-108.

COMPARISON OF MANUAL AND AUTOMATIC SUMMARISATION

VICTORIA ALEXEI¹, VICTORIA BOBICEV¹

¹*Technical University of Moldova*

MD2004, Chişinău, Ştefan cel Mare bd. 168 str., RM

`{victoria.alexei,victoria.bobicev}@ia.utm.md`

Abstract

The task of summarisation has been traditionally a difficult one for the natural language processing field. The new tools based on large language models such as GPT neural net deal with most text processing tasks including summarisation, with remarkable success. However, the modern models are still incapable of text understanding and sometimes this leads to erroneous results, especially when the volume of training data is comparatively small. Initially, works on text summarisation have focused primarily on high resource languages, like English, mostly due to the limited availability of datasets for other ones. GPT-4 is generating automatic summaries for Romanian texts as well; however, there is a need to investigate and evaluate large language models performance on various tasks including such difficult one as summarisation in special in Romanian. In this paper, we experimented with summarisation evaluation. We used Romanian newspaper articles as a dataset. Given the fact that there are already many automatic summarisation instruments, we compared manual and automatic summaries. Two manual and two automatic summaries have been created for each article. ChatGPT and Google BARD automatic summarising tools have been used to create summaries. These summaries were then checked by human experts for correctness, quality, and ability to capture the main ideas of the original text. The ROUGE, BLEU and METEOR metrics were also used for the comparison of different summaries and their quality. We compared manual and automatic summaries created by different tools.

1 Introduction

In modern society, the amount of information a person is processing daily is not only huge, but it is growing constantly; the problem of its reduction is essential. Fortunately, brand new and quite powerful tools for text processing appeared recently to help us cope with this information analysis task. The new tools such

as GPT (Generative Pretrained Transformers) models deal with the most text processing tasks with remarkable success.

The task of summarisation has been traditionally a difficult one for the natural language processing field. In essence, it is a so-called AI-complete task (Shapiro, 1992), the one that requires text understanding to reach a high-quality result. Due to the fact that machine learning models cannot understand texts, extractive and abstractive summarisation methods have been developed in the natural language processing field. The newly created Artificial Intelligence based applications implement more difficult and effective abstractive methods and in many cases with good quality.

However, the modern models are still incapable of text understanding and sometimes this leads to erroneous results, especially when the volume of training data is comparatively small. Initially, works on text summarisation have focused primarily on high resource languages like English, mostly due to the limited availability of datasets for other ones.

In the domain of natural language processing, Romanian language is considered a low resource language (Gifu, 2016) although modern language models, such as GPT-4, are trained on all Wikipedia languages including the Romanian part of it.

GPT-4 is generating summaries for Romanian texts as well, however, there is a need to investigate and evaluate large language models performance on various tasks including such a difficult one as summarisation in special in Romanian.

Evaluation of automated summarisation results is a complex and multidimensional task. The generated text can be evaluated for fluency, coherence, relevance, and many other aspects of interest (Jain *et al.*, 2023). In general, the gold-standard method for evaluating such texts is still human evaluation but this is a costly and time-consuming method.

In this paper, we report our summarisation evaluation experiments. We used Romanian newspaper articles as a dataset; two manual and two automated summaries have been created for each article. Given the fact that the novel automatic summarisation instruments can produce summaries at the level of humans, we compared manual and automatic summaries.

The main research questions are: Are modern tools as good at creating summaries as humans? What is the difference between manual and automatic summaries created by modern tools?

We used ROUGE, BLEU and METEOR metrics for the comparison of different summaries.

The rest of the paper is organized as follows. The next section 2 discusses the related work in document summarisation, modern text generative tools and their use for the text summarisation, advances and problems in summarisation for languages other than English. The last part of this section is dedicated to

summarisation evaluation methods. Section 3 presents the methodology of our experiments; section 4 demonstrates and explores the obtained results, and the last section 5 concludes the paper.

2 Related Work

Automatic summarisation is one of the oldest topics in computational linguistics. The final goal of the research in this domain was to develop a technique, which generates a summary that matches with the human made summary. The first problem here is the goal itself as the human made summary can be quite different and even an assessment of the human generated summary is a difficult task. Automatic summary evaluation is another challenging issue in this research field. The modern natural language generating tools are able to produce rather good summaries although the quality of their summaries has not been thoroughly investigated yet.

2.1 Automatic Summarisation

Automatic text summarisation system generates a summary, i.e. short length text that includes all the important information of the document (Gambhir and Gupta, 2017). The initially sophisticated definitions of a summary as “an abbreviated, accurate representation of a document which should be published with it and which is also useful in secondary publications and services” or “an abstract summarises the essential contents of a particular knowledge record, and it is a true surrogate of the document” have been replaced with a simpler definition such as “a concise representation of a document’s content to enable the reader to determine its relevance to a specific information” (Orăsan, 2019).

Later, a whole taxonomy of summary types has been developed in dependence of goals, methods, and sources of produced summaries. The taxonomy included generic- vs user-focused summaries, user-focused summaries, query-based summaries, single-document vs multi-document summaries; indicative summaries and informative summaries; monolingual, cross-lingual and multilingual summaries (Rush, *et al.*, 2015).

However, there were always only two general methods of automatic summarisation: extractive and abstractive. Extractive summarisation is done by identifying the most informative fragments of the original text and presenting them as they are, often in the order they appear in the source. Abstractive summarisation is more complex than extractive; it generates novel text units that are not present in the original document to convey the most important information from the source. It is obvious that the latter method is computationally more

challenging than extractive methods, but it is more similar to the human’s way to generate summaries.

Obviously, most studies have focused on extractive summarisation as the simpler one. Lately, however, deep learning techniques based on pre-trained large language models (LLM) have been obtaining remarkably good results creating abstractive summaries. Plenty of works presented various automatic summarisation methods and various brief and more wide-ranged overviews have been published, some of the most recent are (Gambhir and Gupta, 2017), (Wafaa *et al.*, 2021), (Allahyari *et al.*, 2017), (Ramya *et al.*, 2023) (Yadav, 2022) describing the approaches, methods, techniques, datasets, evaluation methods, and future research directions of this field.

2.2 Modern Summarisation Tools

Recently, the use of deep learning technologies has led to more accurate results in various fields of computational linguistics including automatic summarisation. Nallapati *et al.* (2016) developed an abstractive text summarisation method using Attentional Encoder-Decoder Recurrent Neural Networks with promising results, outperforming state-of-the-art for that period. Rekabdar *et al.* (2019) used generative adversarial networks in abstractive text summarisation tasks with a time-decay attention mechanism obtaining high ROUGE score in their experiments.

The most difficult challenge in using deep learning is the need for large datasets. After the large pre-trained models have been created, most deep learning methods started to use them intensely. Zhang *et al.* (2020) demonstrated the effects of the pre-training corpora achieving state-of-the-art performance on 12 datasets. However, the computational complexity of BERT-based pre-trained models grows quadratic with the input length, making them impractical for long, or even medium sized text documents. (Salchner *et al.*, 2022) presented Graph Neural Networks demonstrating their advantages in use for text summarisation. Cao (2022) contains a comprehensive discussion of neural network-based summarisation systems and their evaluation problems.

However, the recent success of new generation LLM like GPT-3 (Generative Pre-trained Transformer) (Brown *et al.*, 2020) resulted in a paradigm shift in NLP research. Goyal *et al.* (2023) compared GPT-3 against various models trained on large summarisation datasets publishing the first systematic study of the impact of prompt-based models on the text summarisation research space, using an Instruct-tuned 175B GPT-3 model (text-davinci002) (Ouyang *et al.*, 2022) as a case study.

2.3 Summarisation for Romanian

Most summarisation studies concentrate on English texts; only a small fraction experimented with other languages. Alcantara *et al.* (2023) highlighted that most pre-trained models are only available for English and with the main purpose to summarise texts in German the authors focused on the German BERT multilingual model.

Languages widely spoken in populous countries like China and India receive more focus. Several broad overviews of the field of automatic text summarisation in different languages and various text summarisation approaches focusing on Indian languages but discussing also Arabic, Chinese and some other languages was presented in (Kumar *et al.*, 2021) (Ganesh *et al.*, 2022). Shliazhko *et al.* (2023) presents GPT architecture trained and tested on 61 low-resourced languages demonstrating the possibility to obtain reasonably good productivity even for comparatively small amounts of training data.

We work with Romanian texts. Romanian language is considered under-resourced in the NLP domain, however, there are several papers dedicated to Romanian language processing including summarisation. Dan Cristea *et al.* presented several works dedicated to summarisation, based on discourse structure theories (Cristea *et al.*, 2005) (Anechitei *et al.*, 2013). Niculescu *et al.* (2022) presented a study of novel pre-trained models created for Romanian applied to the summarisation task. Their study concentrated on developing a model for abstractive summarisation in Romanian and creating a dataset for summarisation evaluation. Multiple models based on the Romanian GPT-2 have been used to generate summaries and these summaries have been evaluated by humans. This evaluation demonstrated the ability of large pre-trained models to tackle any language they are trained in. Buzea *et al.* (2022) experimented on training GPT-2 on Romanian texts and demonstrated its ability to generate coherent texts although this type of models requires huge volumes of training data for an acceptable performance.

2.4 Summarisation Evaluation Methods

The task of automatic summarisation evaluation is very difficult. The main problem is the fact that the notions of a good summary are highly subjective. The quality of a summary is determined by such vague characteristics as presence of the important information from a document, absence of less important or redundant information, its coherence and consistency. It is obvious that if several people write a summary from a text, each summary would be different, even if all can be labelled as “correct summary”.

Multiple series of evaluation campaigns and conferences included document summarisation tracks. DUC¹ (Document Understanding Conference 2001-2007) had the most important impact on the field creating evaluation datasets with human-written reference summaries and established the ROUGE (ISI's Recall-Oriented Understudy for Gisting Evaluation) automatic (n-gram) matching (Lin, 2004).

ROUGE-1 counts recall based on matching unigrams, ROUGE-2 counts recall based on matching bigrams, etc. For any n , the total number of n -grams across the reference is calculated and the fraction of them in the candidate summary is the required metric value.

ROUGE-L is based on longest common subsequence (LCS), ROUGE-W is weighted LCS, ROUGE-S is and skip-bigram co-occurrence statistics, respectively. These metrics use an F-score, the harmonic mean of precision and recall values calculated as for ROUGE-L.

Obviously, this metric is more or less appropriate for extractive summaries and much less for abstractive ones, which can be written in other words than in original text and convey the main idea of the original text, having minimum overlap with it.

Another overlap metric is BLEU (Papineni *et al.*, 2002) actively used in machine translation and in summarisation. Blagec *et al.* (2021) pointed out that ROUGE and BLEU are still the most often used metrics and discussed their weaknesses. METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Denkowski and Lavie, 2014) is a metric used to evaluate the quality of machine translations and to improve certain limitations of BLEU and ROUGE. METEOR calculates weighted F-score based on mapping unigrams. Precision and recall are calculated as m/c and m/r , where m is the matches' length, and c and r are candidate and reference lengths, respectively. METEOR also uses a penalty in its formula; the penalty decreases the score if the word order is broken (Satanjeev and Lavie, 2005).

The most straightforward way to evaluate a summary is to have a human assess its quality. The factors that human experts must consider when giving scores to each candidate summary are grammaticality, non-redundancy, integration of most important pieces of information, structure, and coherence (Allahyari *et al.*, 2017). The drawback of this method is the necessity of human work, which is costly, time consuming and error prone. For example, Clark *et al.* (2021) demonstrated that human evaluators need detailed instructions with examples and some training to be able to perform a qualitative evaluation of generated text.

¹ <https://duc.nist.gov/>

3 Methodology of the Experiments

The main research questions of our study are: Are modern tools as good at creating summaries as humans? What is the difference between manual and automatic summaries created by modern tools?

In order to answer this question, we:

- 1) Calculated the values of classical metrics for manual summaries created by humans.
- 2) Calculated the values of classical metrics for automatic summaries.
- 3) Calculated the difference between manual and automatic summaries measured by classical metrics.
- 4) Manually assessed and compared the quality of manual and automatic summaries.

For our experiments, we selected articles from *Evenimentul zilei* newspaper, years 2001-2003. The corpus has been offered by Constantin Orasan². Taking into consideration that we plan to create manual summaries, we selected 30 short articles with the number of words between 600 and 1000. The selected texts are predominantly about political and social events.

First, we organised our students to write summaries for these articles. The instructions for the task were formulated as follows. “Write a summary for the text. It should be created using extractive methodology, you should select the most important fragments from the original text and organise them in a coherent summary connecting in the same order that they were in the original text. The length of the summary should be exactly 1/3 of the original text measured in words.” In this way we obtained two manually created summaries for each article.

Second, we created two automatic summaries. Currently, there are multiple tools that can create summaries (Ouatu and Gifu, 2020). We analysed a dozen various AI (Artificial Intelligence) text generative tools such as ChatGPT, Microsoft Bing, YouChat, Claude, Jasper, Perplexity AI, LLaMA, Google BARD and some others. We checked if they are free to use, languages they are working with, what interface they have, what are their advantages and drawbacks. Finally, we decided to use two well-known tools: ChatGPT and BARD. For each article, we created two summaries, one with ChatGPT and one with BARD. The prompts had been formulated as follows: “create an extractive summary of 30% of the following text.”

As a result, we obtained four summaries for each article, two written by students and two created by AI tools. The next step is their comparison. It includes:

² <https://www.surrey.ac.uk/people/constantin-orasan>

- calculation of classical metrics for summarisation evaluation: BLUE, ROUGE and METEOR for each summary;
- comparison of two summaries created by AI tools to explore their similarities;
- assessment of all four summaries by students for their relevance and coherence;
- comparison of the best summary written by students with the best one created by AI tool to see how similar they are.

The comparisons have been done by calculating the BLUE, ROUGE and METEOR scores as measures of similarity.

The manual assessment was organised as follows. Each student received a folder with five files: the original article and four summaries. The source of summaries was hidden; evaluators had no knowledge which summary was written by the tool and which by a person although they did know that two summaries were written by humans and two by text-generating applications. The evaluator's task was to evaluate each summary by scoring two characteristics: **relevance** and **coherence** with a score from 1 to 5. The characteristics have been described in more details: Relevance - How well the summary presents key points of the article? Is the information provided by the summary consistent with that in the article? Coherence – How well written, coherent and readable is the summary?

4 Obtained Results and Discussion

The results of the first part of our experiments are four summaries for each of 30 selected articles. The first two have been written by students and were intended to be the so-called “gold standard” to compare automatic summaries with. However, the quality of summaries varied as much as students' capabilities; they have created summaries within their potential. Some summaries were longer than 1/3 of the text; in some cases, the order of sentences in the summary was different in comparison to the article, mixing up the order of ideas of the article.

The average length measured in words for the initial articles and the manual and automatic summaries are shown in Table 1. It yields that students respected the requirement to make the summary exactly 1/3 of the initial text; AI tools, however, being asked the same thing, made shorter summaries.

Several measures have been used to evaluate the performance of automatic summarisation. BLEU 1 precision is defined as the number of words in the automated summary that are also present in the reference summary divided by the number of words in the summary. BLEU 1 score is calculated by multiplication of the Precision Score with the Brevity Penalty. The brevity penalty is calculated

as the ratio of the length of the resulting text to the length of the reference text. If the resulting text is shorter than the reference text, the brevity penalty decreases.

Table 1. Average volume of initial article, average volume of summaries in percents.

Type of Text	Volume
Average volume of initial articles (in words)	733
Average volume of manual summaries in percents	35%
Average volume of ChatGPT summaries in percents	23%
Average volume of BARD summaries in percents	22%

METEOR: a measure of similarity between an automated summary and a reference summary that considers both maximum common length of subsequences and word frequency.

ROUGE 1 is based on the frequency of occurrence of common words in the summary and in the reference. ROUGE LSum measures the frequency taking into account word order. A higher ROUGE LSum value indicates that the summary is more accurate and relevant, as well as more coherent.

Table 2 contains average similarity metrics calculated for all four summaries. The first three lines in the table contain scores calculated for summaries against the original texts. The values for the row **Summaries students** are calculated as the average for all **BLEU 1 scores** for all students' summaries; the row **Summaries ChatGPT** is calculated as the average for all **BLEU 1 scores** for all summaries created by ChatGPT; **Summaries BARD** contains the same information for summaries created by BARD.

The last two rows of Table 2 contain values of similarities between students' summaries and ChatGPT (**Students-ChatGPT**) and between BARD and ChatGPT summaries (**BARD-ChatGPT**).

Table 2. All calculated metrics for summaries created by students, ChatGPT and BARD (first three rows) and metrics calculated for comparison between summaries created by students and created by ChatGPT and comparison between ChatGPT and BARD.

Type of Text	BLEU 1 score	BLEU 1 precision	Meteor	ROUGE 1 score	ROUGE Lsum
Summaries students	0.128	0.831	0.240	0.423	0.346
Summaries ChatGPT	0.021	0.479	0.068	0.182	0.114
Summaries BARD	0.024	0.479	0.067	0.184	0.127
Students-ChatGPT	0.190	0.342	0.145	0.277	0.168
BARD-ChatGPT	0.288	0.366	0.051	0.025	0.248

For the first three rows in Table 2 with summaries against initial texts, BLEU 1 precisions for students' summaries are comparatively high; this confirms that students made extractive summaries and used more original words from initial texts compared to summaries created by ChatGPT and BARD. BLEU 1 scores are lower than precision because of the brevity penalty for the summary; but the same trend is maintained: students' summaries have higher scores compared to BARD and ChatGPT. The same tendency is seen for all used metrics; students wrote summaries that were more similar to the original text than AI tools. METEOR was designed to be more robust to linguistic and cultural differences, being able to handle wider variations of expressions and grammatical structures, while BLEU tends to be more rigid and place more weight on identical words or n-grams. However, Meteor score is lower; probably, the penalty for word order influenced the result. This suggests that both students and AI tools reordered texts while summarising.

ROUGE was designed particularly for summarisation evaluation thus lacking the problems of BLEU. However, its score is still low for AI tools; for students ROUGE score is similar with other studies. It is interesting that the scores for ChatGPT and BARD are so close, almost identical for all metrics.

The last two rows of the table present similarity scores between summaries written by students and AI tools. We see that the similarity between ChatGPT and BARD summaries measured by BLUE is higher than similarity between students and ChatGPT. However, METEOR and ROUGE 1 score did not find this similarity between ChatGPT and BARD although ROUGE LSum also demonstrates a high value of similarity between AI tools. It is interesting why ROUGE 1 and ROUGE LSum behave so differently.

ROUGE 1 score maintains the same tendency as Meteor score. Summaries students are higher than summaries GPT and BARD.

It is worth mentioning that METEOR results are better for Summaries students and for Students-ChatGPT, in comparison to BLEU score, because it considers word order in its evaluation, comparing not only words but also their order in the sentence.

Boxplots in Figures 1 and 2 clearly demonstrate the difference between students' summaries and AI tools' results. All metrics are considerably lower for the AI tools. It is explainable, the metrics are working well with extractive types of summaries and the tools are producing abstractive ones. They actively paraphrase and use synonyms. Students were instructed to create extractive summaries, extracting fragments of the initial text, and connecting them to create a coherent summary. As it is seen from the plots, only a small part of the students followed the instructions; they mostly rewrote the initial text thus obtaining low similarity scores.

COMPARISON OF MANUAL AND AUTOMATIC SUMMARISATION

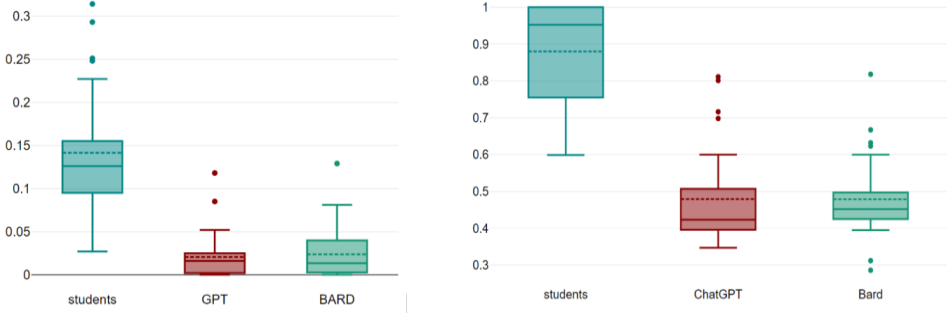


Figure 1. Boxplot for the average BLEU 1 score (left) and BLEU 1 Precision (right) for summaries created by students, ChatGPT and BARD.

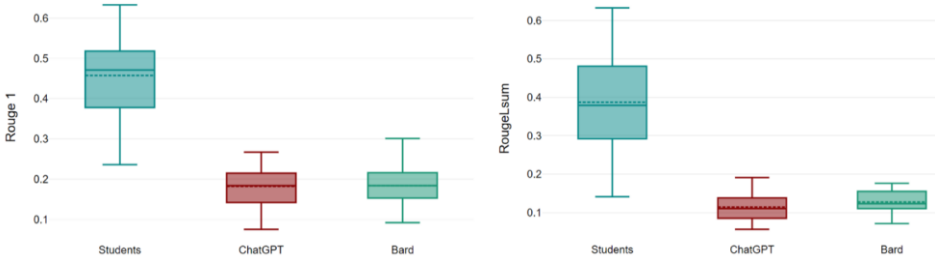


Figure 2. Boxplot for the average ROUGE 1 score (left) and ROUGE Lsum (right) for summaries created by students, ChatGPT and BARD.

Table 3 and Figure 3 present the average results of students' manual evaluations of the summaries for two characteristics: relevance and coherence. We can observe that all values are quite similar, they vary around 3 and 4 for all types of summaries and both metrics. We assume that students tended to avoid the extreme grades 1 and 5 mostly because of psychological reasons. Grading the summaries, students were not aware of which summaries had been created by their colleagues and which by AI tools and trying to guess while in the process of grading they have difficulties to do so. Based on the results, we can say that in general, AI tools are slightly better. Their summaries on average were a little more relevant and more coherent, in special, BARD.

Taking into consideration that all metrics used for rating used solely words and their combinations we understand that they cannot evaluate properly abstractive summaries, which are written in different words and phrases.

Two independent persons have graded each summary; the last two columns in Table 3 contain average differences between the ratings. The inter-rater agreement is extremely low as there are 5 possible variants for grading. In most cases, there were differences in 1 point, for example, 3 and 4 or 4 and 5 but bigger differences

are present as well. It is understandable, a summary is a result of subjective action, and its evaluation as well is quite subjective. In some cases, raters had difficulties understanding the initial text; in some cases, they considered that the summary had to be written differently. Plenty of reasons could influence the raters' decisions.

Table 3. Average relevance and coherence for all summaries assessed by students manually by rating them from 1 to 5 and average difference between two graders.

Type of summary	Relevance	Coherence	Difference for Relevance	Difference for Coherence
Summaries students 1	3.68	3.62	1.17	1.03
Summaries students 2	3.77	3.55	1.00	1.17
Summaries ChatGPT	3.75	3.77	1.30	1.07
Summaries BARD	3.38	3.58	1.37	1.17

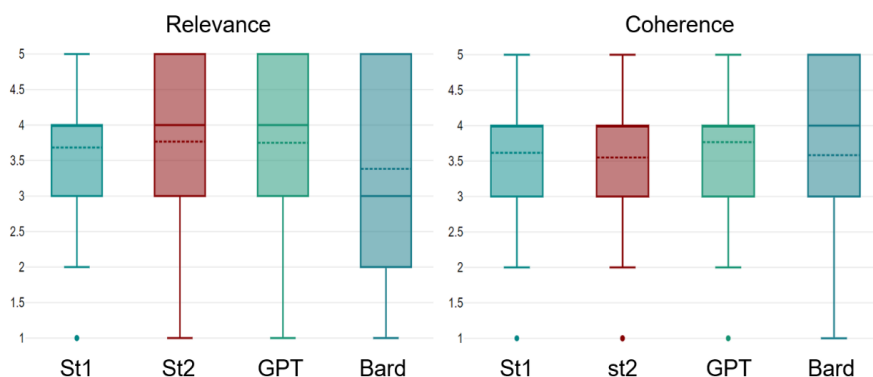


Figure 3. Boxplot for relevance (left) and coherence (right) for all summaries assessed by students manually by rating them from 1 to 5. St1 and St2 refers to the summary created by one student and by another student as we obtained two summaries written by students for each text.

5 Conclusion

This paper presents experiments on summarisation evaluation. Two manual and two automated summaries have been created for 30 Romanian newspaper articles. Given the fact that there are already many AI-based automatic summarisation instruments, we compared manual and automatic summaries written by our students and created by ChatGPT and Google BARD. The classical metrics BLEU, ROUGE and METEOR have been calculated for the comparison of different summaries and their quality. We found that in our case their values were

rather low as they calculate the rate of overlapping words and their combinations, but both students and AI tools created abstractive summaries using words and phrases that differ from the original text.

We also asked independent raters to assess the summaries quality, namely, their relevance to original text and their coherence. The main conclusion of this experiment is that raters could not tell which summaries were written by students and which were created by AI tools, and rated the summaries created by tools slightly better. However, the inter-rater agreement was extremely low and more experiments should be done to guarantee the reliability of the obtained results.

We plan to collect more summaries created by people and tools and organise more experiments for better comparison of human- and tool-made texts.

One more conclusion can be made based on our experiments: Modern AI-based applications are capable of manipulating Romanian texts creating sufficiently good results.

References

- Alcantara, T. H. M., Krütli, D., Ravada, R., and Hanne, T. 2023. Multilingual Text Summarization for German Texts Using Transformer Models. *Information* 14, no. 6:303. <https://doi.org/10.3390/info14060303>
- Anechitei, D., Cristea, D., Dimosthenis, I., Ignat, E., Karagiozov, D., Koeva, S., Kopeć, M., and Vertan, C. 2013. Summarizing Short Texts through a Discourse-Centered Approach in a Multilingual Context. In *Where Humans Meet Machines: Innovative Solutions to Knotty Natural Language Problems*, Neustein, A., Markowitz, J.A. (eds.). Springer Verlag, Heidelberg/New York, pages 109-136.
- Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E., Gutiérrez, J., and Kochut, K. 2017. Text Summarization Techniques: A Brief Survey. *International Journal of Advanced Computer Science and Applications (IJACSA)*. 8:397-405. 10.14569/IJACSA.2017.081052.
- Banerjee, S. and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan. Association for Computational Linguistics. pages 65-72.
- Blagec, K., Dorffner, G., Moradi, M., Ott, S., and Samwald, M. 2022. A Global Analysis of Metrics Used for Measuring Performance in Natural Language Processing. *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pages 52-63.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. 2020. Language Models are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877-1901.
- Buzea, M. C., Trăușan-Matu, Ș., and Rebedea, T. 2022. Automatic Romanian Text Generation Using GPT-2. *U.P.B. Sci. Bull., Series C*, 84(4):15-30

- Cao, M. 2022. A Survey on Neural Abstractive Summarization Methods and Factual Consistency of Summarization. *ArXiv, Computer Science – Computation and Language*. <https://arxiv.org/abs/2204.09519>
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. 2021. All That’s ‘Human’ Is Not Gold: Evaluating Human Evaluation of Generated Text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pages 7282-7296.
- Cristea, D., Postolache, O., and Pistol, I. C. 2005. Summarisation through Discourse Structure. In *Computational Linguistics and Intelligent Text Processing*, Alexander Gelbukh (ed.), 6th International Conference CICLing 2005, Mexico, Proceedings Springer LNSC, 3406:632-644.
- Denkowski, M. and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.
- El-Kassas, W. S., Salama, C. R., Rafea, A. A., and Mohamed, H. K. 2021. Automatic Text Summarization: A Comprehensive Survey, *Expert Systems with Applications*, Volume 165, 2021.
- Gambhir, M. and Gupta, V. 2017. Recent Automatic Text Summarization Techniques: a Survey. *Artif. Intell. Rev.* 47(1):1-66. <https://doi.org/10.1007/s10462-016-9475-9>
- Ganesh, A., Jaya, A., and Sunitha, C. 2022. An Overview of Semantic Based Document Summarization in Different Languages, *ECS Transactions*, Volume 107, Number 1.
- Gifu, D. 2016. Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language. PhD thesis, “Alexandru Ioan Cuza” Univ. Iași, Iași, Romania.
- Goyal, T., Li, J. J., and Durrett, G. 2022. News Summarization and Evaluation in the Era of GPT-3. *ArXiv:2209.12356*. <https://arxiv.org/abs/2209.12356>
- Jain, S., Keshava, V., Sathyendra, S. M., Fernandes, P., Liu, P., Neubig, G., and Zhou, C. 2023. Multi-Dimensional Evaluation of Text Summarization with In-Context Learning. *ACL 2023*, pages 8487-8495.
- Kumar, Y., Kaur, K., and Kaur, S. 2021. Study of Automatic Text Summarization Approaches in Different Languages. *Artificial Intelligence Review.* 54. 10.1007/s10462-021-09964-4.
- Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out. Association for Computational Linguistics*, Barcelona, Spain. pages 74-81.
- Nallapati, R., Zhou, B., dos Santos, C. N., Gulcehre, C., and Xiang, B. 2016. Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280-290.
- Niculescu, M. A., Ruseti, Ș., and Dascalu, M. 2022. RoSummary: Control Tokens for Romanian News Summarization. *Algorithms*; 15(12):472. Available online at <https://doi.org/10.3390/a15120472>
- Orăsan, C. 2019. Automatic Summarisation: 25 years on. *Natural Language Engineering*, 25(6), 735-751. doi:10.1017/S1351324919000524

- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., *et al.* 2022. Training Language Models to Follow Instructions with Human Feedback. *arXiv preprint* arXiv:2203.02155.
- Papineni, K., Roukos, S., Ward, T., and, Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311-318.
- Ramya R. S., Dayananda Sagar, M. Shahina Parveen, Savitha Hiremath, Isha Pugalia, S. H. Manjula, and K.R. Venugopal. 2023. A Survey on Automatic Text Summarization and its Techniques. *International Journal of Intelligent Systems and Applications in Engineering*, 11(1s), pages 63-71. Available and retrieved from <https://ijisae.org/index.php/IJISAE/article/view/2478>
- Rekabdar, B., Mousas, C., and Gupta, B. 2019. Generative Adversarial Network with Policy Gradient for Text Summarization, *IEEE 13th International Conference on Semantic Computing (ICSC)*, Newport Beach, CA, USA, pages 204-207, doi: 10.1109/ICOSC.2019.8665583.
- Ouat, B. and Gifu, D. 2020. Chatbot, the Future of Learning? *Proceedings of the 5th International Conference on Smart Learning Ecosystems and Regional Development (SLERD 2020)*, in Ludic, Co-design and Tools Supporting Smart Learning Ecosystems and Smart Education, Springer, 2020, pages 263-268.
- Rush, A. M., Chopra, S., and Weston, J. 2015. A Neural Attention Model for Abstractive Sentence Summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal. Association for Computational Linguistics. pages 379-389.
- Salchner, M. F. and Jatowt, A. 2022. A Survey of Automatic Text Summarization using Graph Neural Networks. *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6139-6150.
- Shapiro, S. C. 1992. *Artificial Intelligence*. Encyclopedia of Artificial Intelligence (Second Edition), *Stuart C. Shapiro (Ed.)*, pages 54-57. New York: John Wiley.
- Shliazhko, O., Fenogenova, A., Tikhonova, M., Mikhailov, V., Kozlova, A., and Shavrina, T. mGPT: Few-Shot Learners Go Multilingual. *arXiv:2204.07580v2*. <https://arxiv.org/abs/2204.07580>
- Yadav, D., Desai, J., and Yadav, A. K. 2022. Automatic Text Summarization Methods: A Comprehensive Review. *ArXiv, abs/2204.01849*. <https://arxiv.org/abs/2204.01849>
- Zhang, J., Zhao, Y., Saleh, M., and Liu, P. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning, Hal Daumé III and Aarti Singh (eds.)*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328-11339.

CHAPTER 4

LANGUAGE MODELS

THE NEWS IN BRIEF – LEVERAGING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE IN NEWS CLUSTERING, SUMMARISATION AND EVALUATION

BOGDAN FLOREA¹, ADRIAN IFTENE¹

¹*“Alexandru Ioan Cuza” University of Iași, Romania, Faculty of Computer Science*

{bogdanvflorea, adiftene}@gmail.com

Abstract

This paper proposes a solution for the problem of news overload and the effort we make to stay connected to contemporary political, economic, and social realities. The amount of information made available to users on the Internet is immeasurable, far exceeding the time and systematisation capacity of a simple person. Using text analysis, entity extraction, K-means clustering, and a text-to-text transformer the proposed solution offers the classification, filtering, and summarisation of large volumes of data uploaded daily in the digital space. Also, the paper pleads for a cross-media platform news comparison, using data provided by different observers, in order to mark out polarisation and ideology in news.

1 Introduction

We live in a time when information channels are changing significantly. The classic means of media communication are declining in terms of audience, with only the older generations remaining loyal to them. Young people get more information from the Internet through news sites and social networks (Mineia and Iftene, 2017). We are witnessing the rise of *social media* and *video streaming* platforms not only as a space for socialising and entertainment but also as a public forum for political debate (Gifu and Cristea, 2012), a platform for the assertion of amateur journalism, a vector for the spread of fake news and even ideologies (Abonizio *et al.*, 2020). On a theoretical level, in the field of news analysis, two research directions have emerged, with remarkable results so far: the linguistic approach and the one based on social networks. We have followed the first approach, adding a comparative nuance to it, focusing on the analysis of

professional news sources, and press institutions active on the internet. What is new in our proposed solution is that we do not limit ourselves to the linguistic analysis of a news story taken in isolation or just in the context of its source. We strive to gather thousands of the latest news articles and pass them through summarisation, entity extraction, clustering, and linguistic analysis to extract the most important topics of the day, and for each topic the most relevant and well-written pieces. We place each piece of information in the broader context of other articles and news sources to see the recurrence of the headline, the different angles from which the event has been viewed, how long the interest of journalists and the public extends for a certain subject, or on the contrary, how certain topics are seen as marginal and are scarcely covered by the media, how certain channels position themselves politically or ideologically, and much more. Thus, our proposed approach is holistic, based on combining various methods from *Machine Learning*, considering principles from journalism ethics (Gîfu and Cioca, 2013), and propelled by the power of Artificial Intelligence. It is relevant and can be deepened in interdisciplinary research at the intersection of Computer Science with Sociology, Journalism, or Political Science.

The implementation of our theoretical solution is a very useful web application that processes a huge amount of recent news, classifies it, groups it by common entities, and filters out the least important topics. In the end, it gives the user the best briefing of the day on politics, technology, science, and sports, with summarised and evaluated content. On a research level, our main goal was to explore the potential of combining clustering methods from the field of *Machine Learning* with the power of *AI* to summarise articles, group them thematically, or analyse them linguistically. The broader objective of this endeavour is to contribute to the promotion of quality news, curb the trend of ideological polarisation in society, cultivate critical thinking among readers (as suggested by Machete and Turpin (2020)), and limit misinformation. In addition, we also have an interdisciplinary objective: to highlight the extraordinary potential that arises from the convergence of fields of knowledge, in our case, between information technology and socio-political sciences.

The paper is divided into three parts, along with an introduction and conclusions. The first part highlights the problem to which we propose our solution: the information explosion and the limited time we have to keep abreast of the major events of the present. Also, we draw attention to some of the existing software solutions for reading news, as well as a couple of interesting initiatives in the journalism field to outline the deontology of online press providers. In the second part, we provide documentation of our application, detailing its architecture, interface, functionalities, APIs, collections, and packages used. In the

last part we discuss the problem of news *clustering* and *summarisation*, we report the experiments conducted and the results obtained.

2 Information Overload and the Need to Keep Pace with the World

The present time can be aptly characterised by (Naisbitt, 1982) observation: “*We are drowning in information, but starving for knowledge*”. It captures a severe issue in contemporary society: the fact that although we have access to more information than at any other point in history, we often fail to translate this information into knowledge or wisdom that positively transforms our lives. Countless news is published every day using radio, TV, and the internet making it virtually impossible for a single person to keep track of everything that happens. Therefore, we face a significant challenge in choosing where we get our information, how much time we spend browsing the news, how we discern between different viewpoints, and how we identify fake news (Cuşmuliuc *et al.*, 2021a), (Cuşmuliuc *et al.*, 2020).

2.1 Limited Time, Unlimited Information

In the field of news or *mass-media* this information explosion has democratised access to information - even in totalitarian regimes people have more and more access to the outside, free world - and has enriched the sources of information for ordinary citizens. Every citizen can be an “amateur journalist” if, for example, they photograph or record an event with their mobile phone and post it on a social network. This ease of spreading information also has its negative aspects: a malevolent or misguided person can release false, distorted, or at least incomplete information into the digital space, which can quickly become viral, causing serious consequences (Valenzuela and Arriagadac, 2014). Another downside is that this explosion of data is overwhelming for people trying to keep up with the world they live in (Bawden and Robinson, 2009). This can become a major time-consuming endeavour if we consider the large number of press providers and the even larger number of articles that appear in print or digital format every day. The total number of news articles published globally each day through all means is impossible to specify precisely because of the large number of news providers and their various formats: online, print, or broadcast. However, it is estimated that daily, between 2 and 3 million journalistic articles are published worldwide (Campbell, 2023). This estimate does not take into account the articles published on *social media* platforms, where the amateur journalism we mentioned is a widespread phenomenon.

In this context, people naturally narrow their areas of interest both geographically and thematically, following only a small selection of news

programs on TV, news sites on the internet, or people on *social media*. Many prefer having a news application installed on their mobile phone to be notified in real-time of recent events. But although we narrow down our areas of interest and sources of information, the whole process may consume several hours every day and the frequent switching of communication media and news sources, along with notifications on our phones, negatively impact our ability to concentrate on work or home tasks, or our loved ones. Having a selective focus is very useful, and especially young people gather all the information they need from *social media* (Atodiresei *et al.*, 2018). However, this attitude raises the issue of the veracity of information, of the trust in their source, of the algorithms that, depending on our user profile, only suggest certain news to us. Some researchers have shown that within social networks the boundary between rumour and news is increasingly blurred (Vosoughi *et al.*, 2018). Another worrying aspect in the context of social networks, highlighted by Soroush Vosoughi in his study, is that messages with false claims are 70% more likely to be re-shared than messages with true stories (Vosoughi *et al.*, 2018). The problem with the algorithms that build our personalised flow of messages, notifications, or videos is they determine which information is more important to us and which is omitted, based on our previous choices, our digital profile shaped over time, and commercial interests promoting certain messages. This reality leads to the danger highlighted by Pariser (2011) in his book *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. We tend to follow more and more people and entities who share similar interests and aspirations to ours, and less and less those who have different opinions. In this way, information that we do not like, but that would be good to know, we may not find out. We end up living in an ideological fort close to the outside. So, because time is limited and information is unlimited, we need to focus. But we cannot focus on a single platform or source of information, except at the risk of narrowing our horizon of knowledge, being misinformed, and slipping into ideological attitudes. To keep our attention on what is important and save time we feel the need for systems or applications that group the news, making a selection from several sources, on different spheres of interest and degrees of importance.

2.2 News Aggregators - a Possible Solution to the Great Influx of News

In this era of rapid information dissemination, the challenge is no longer the availability of news, but rather its immeasurable volume. We could spend hours scrolling through the news and still only scratch the surface of the information universe created by a multi-polar, dynamic, and interconnected world at a level unprecedented in human history. In this context, applications that make a magazine of the press and group the news have emerged as a necessary tool for

those who want to be informed. These platforms aggregate articles from various sources into a single interface, allowing users to navigate the wide range of news available.

*Inoreader*¹ application is developed in Eastern Europe, in Bulgaria, and offers in the browser or on mobile the tool to search news by keywords, the possibility to save favourite articles, to translate articles from one language to another, to pause reading, and to pick up where you left off, to define news sources for your personal feed, to monitor keywords across news stories, to apply actions - such as tagging or generating notifications - to articles based on user-defined rules, to apply filters to the news feed, and to highlight words or phrases within articles. *Feedly*² proposes that the user set their main news sources and collaborate with an artificial intelligence engine called Leo, which trains to distinguish between truly important news and those categorised as less important or simply “noise”.

*Pocket*³ is a free service for saving, grouping, and exploring articles from various fields. It is available on multiple devices and operating systems, and some functionality is offered on a subscription basis. The usage mode allows the declaration of favourite news sources and topics of interest. From these, Pocket creates a personalised feed of news and articles. The user’s profile is updated based on in-app activity, with recommendations in step with areas of interest. An article can have written or media content and can remember where you left off in the event of an interruption in its flow. Also, while surfing the Internet you can save articles of interest to browse them later using the Pocket extension for the browser.

*Flipboard*⁴ is an application available on *Windows*, *Android* or *iOS* operating systems. It brings together in one place the user’s favourite sources of information for fields ranging from politics to technology or travel. *Flipboard* aims to bring together a plurality of perspectives, quality content, and the most important news of the day. It gives you the possibility to create your newspaper where you will find daily news from sources and fields that interest you. *Google News*, from the tech giant *Google*, offers users a web page dedicated to news, as well as a similar service available on Android and iOS. Depending on the device’s location, the user’s profile, and their express preferences, *Google News* gathers in one place a very wide range of journalistic materials from major international news agencies, local media outlets, and even blogs and magazines. We can group some favourite sources of information based on which the application will provide us with the

¹ <https://www.inoreader.com/>

² <https://feedly.com/>

³ <https://getpocket.com/en/>

⁴ <https://flipboard.com/>

main news several times a day. If we use the *Google voice* assistant, news can be listened to by a voice command even when we drive the car or do other activities that occupy our hands and eyes. When the user's attention is directed to an important event, *Google News* can provide a top of the news about that topic, allowing it to be evaluated from several points of view. The application also records the activity in our profile and generates a section of recommended news for us. Those from *Apple* also offer an application dedicated to news: *Apple News*. The biggest news agencies have their web services and mobile apps, notable here being *Associated Press* or *AP News*. They do not group news from different sources, but they have the ability, through the size of their organisation, to provide quality news in most fields and from almost all meridians of the world.

It is useful to read the news in one app, but we can also have some clues about the quality of the news we're browsing. This is with the help of extensions installed in our browser that attach to each article or news source a score that is intended to reflect their journalistic quality. One such tool is *NewsGuard*⁵ which aims to combat disinformation and increase the quality of news in the digital space. Behind the app is a team of journalists who analyse and rate online news sources according to several criteria, such as the prevalence of fake content, transparency about funding and patronage, or clear labelling of advertising. All ratings are done by human analysts, and are reasoned, and when we use the web browser extension each news site will have a credibility and transparency score attached. In this way *NewsGuard* Technologies hopes to combat disinformation and promote high-quality media.

Our research aims to achieve the grouping of news from multiple sites, agencies, or television stations based on events, personalities, geographic locations, or other entities that they address in common. This clustering would give us a broad view of the events because it would bring together different perspectives, and on-the-spot testimonies from various angles and could lead us to the desirable balanced and comprehensive understanding of the available information (Steinbach *et al.*, 2000). Once gathered in clusters that have the same topic, the challenge remains to go through the journalistic material which can be very long. Thus, a need to filter out the least important clusters and to summarise, to highlight the main points of each article surrounding the big headlines (Nenkova and McKeown, 2011). This leads to the efficiency of information consumption: the ability to know the key points of the main events and debates, in a shorter time, remaining at our discretion to read *in extenso* those articles that particularly concern us. None of the reviewed apps make use of abstractive summarisation and only Google News offers a top of the news on a certain topic,

⁵ <https://www.newsguardtech.com/solutions/newsguard>

but not a top of the main topics (meaning the most written about events) in a certain region (like EU, USA) or domain (like politics, economy), with their subsequent and relevant articles. The last beneficial aspect of grouping and summarising news we want to highlight is of an academic or scientific nature. As researchers, it helps us identify trends and patterns of news reporting. We can determine certain ideological affiliations of some media institutions. We can understand the evolution of public attitudes towards a social reality. In the field of economic news, specialists can make forecasts about the financial course of some companies or countries.

3 The News in Brief - Proposed Solution

We implemented an application that enables users to access in a single digital space hundreds and even thousands of recent news from all meridians. The news is divided into geographical and domain categories and is grouped thematically based on the same event or common entities - people, organisations, and places (similar with Cuşmuliuc *et al.* (2021b)). Each story has a summary, a list of keywords, and a sentiment and objectivity rating. The application targets people who mainly get information from the online environment, daily or occasionally, and especially those who want to form an overall, equidistant view of the world we live in, in a reasonable time. *The News in Brief* is a web application, with a basic server-client architecture as seen in Figure 1, where the back-end part is a suite of original or imported methods, written in the python language. It uses several machine learning algorithms and a transformer model (*T5 or Text-To-Text Transfer Transformer*) for text classification and summarisation, as well as some external *APIs* to translate text and extract entities from it. An *SQLite* database stores press articles. On the *front-end* side, users interact with a single-page application developed in *TypeScript* language and supported by the *Angular* framework.

The server is based on the web framework Flask and communicates with the client side through HTTP requests and responses, querying the database. The relational *SQLite* database is a lightweight choice, made for easy integration and fast communication with the application. It stores press articles, their metadata, clusters, and news sources. The server starts an agent (coroutine) that periodically updates the database. This type of asynchronous programming (concurrency) is also used in other contexts in the application for improved performance. The agent relies on both a neural network, a Text-To-Text Transfer Transformer, and several external *APIs* and python collections in processing news articles.

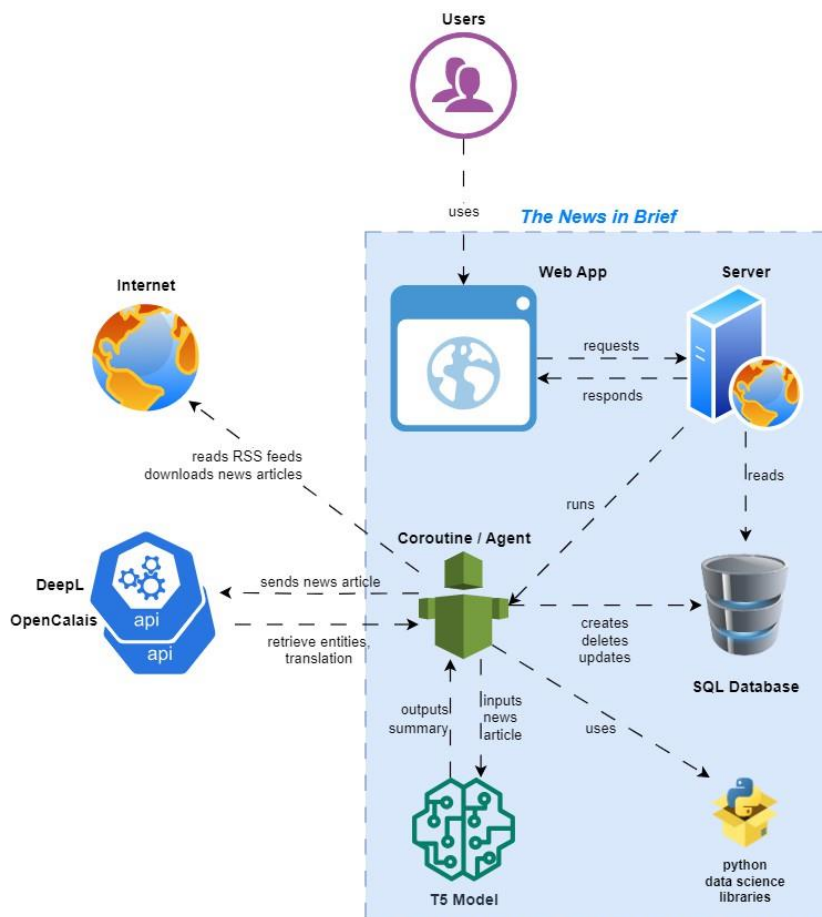


Figure 1. Architecture and components.

The application interface design was guided by the simplicity of navigation, clarity of the sections, attention capture through images, and interactivity through visual feedback. We used elements from the Angular Material collection to provide a clean, unified, and aesthetically appealing look, as well as ease of navigation and efficiency in rendering information. From a chromatic point of view, we chose a palette of shades of blue to indigo, due to the feeling of calmness and professionalism that these colours evoke (see Figure 2). *The News in Brief* offers functionalities to make users effort to find information more efficiently and at the same time to make this a pleasant and easy experience.

Clustering. The application divides the news articles from the database into four geographic categories: world, USA, Europe, Romania. Within each category, the articles are grouped thematically with the help of K-means algorithm from the

field of *Machine Learning*. The groups are filtered according to the number of articles from distinct sources they contain, sorted chronologically, and displayed to the user. Clusters represent headlines and are filtered out when they have few or old elements inside. Thus, marginal news is ignored, and the big headlines of the day or week are played, reducing the information load from hundreds or thousands of articles to only a few dozen.

Abstractive summarisation. Each news article in the app is accompanied by an abstract summary, giving users a concise overview of the article’s main points. This feature, along with the list of entities or keywords, allows users to quickly understand the information without having to read the entire article.

Classification by areas. *The News in Brief* uses a Multinomial Naive Bayes classifier to categorise articles into different interest areas such as politics, technology, science, sports, and others. Using these search tags, users can browse articles by specific domains of interest according to their preferences, without the effort of searching for them among dozens of other news. For example, by typing “technology” in the search bar you get the latest news in technology from all over the world.

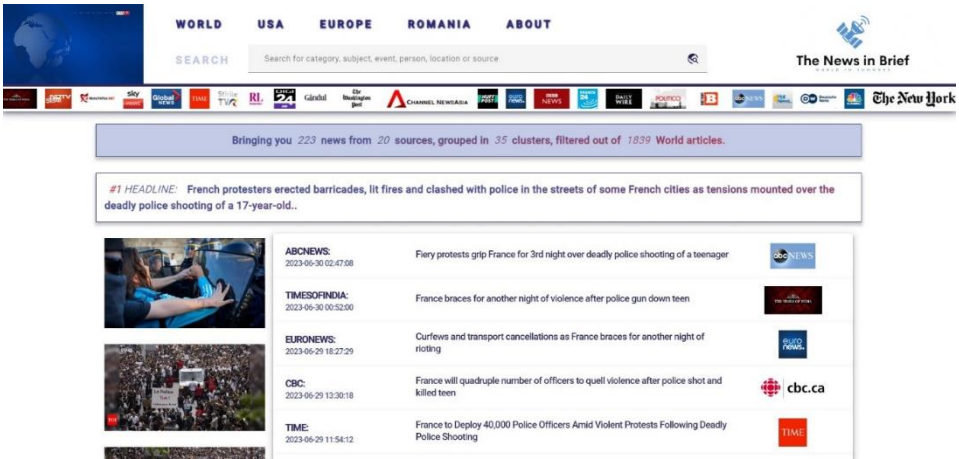


Figure 2. Application overview.

Sentiment analysis, objectivity, and neutrality. When selecting an article, the application performs a sentiment analysis using a lexicon and a rule-based approach to provide an understanding of the emotional tone conveyed. With the use of a trained Naive Bayes classifier, a subjectivity/objectivity score is obtained, so readers can assess whether the article contains accurate statements, precise data, or more personal opinions and approximations. And for a double check a trained Linear Support Vector Classifier calculates the percentage of positive versus negative sentences in the article. Statistics are provided in the form of

charts over these numbers. This feature allows users to quickly evaluate the register and quality of a news story, providing additional information beyond the textual content. If needed, articles with a low neutrality or objectivity score could be sorted out from the start. With this approach we intended to overpass the classical binary fake-real classification (as pointed out by Meinert *et al.*, 2018) and to provide further information to the users, leaving up to them to decide what's worth reading.

Search function allows users to search for specific articles or topics. Users can enter keywords, events, names of people, locations, institutions, or entire sentences to retrieve relevant news articles that match their search data. This functionality allows users to quickly find articles of interest, even among the initially filtered, marginal, less written-about news, but which match the current query.

The News in Brief uses mostly RSS feeds from news agencies from around the globe. However, from the multitude of available sources, we have manually selected approximately 40 national, European, or global press agencies and websites. Some of the criteria that were the basis of the selection are: free access to the news, the prestige of the organisation, the broad observance of the principles of good practice enunciated in The Trust Project (Lehrman, 2023), the frequency of update for the RSS feed and the consistency between the title and content (no click-bait words usage). Here is a selective list of news sources used: Digi24, Gândul, Euronews, BBC, The New York Times, The Guardian, Fox News, Al Jazeera, Times of India, CNBC, Washington Post, ABC News, NDTV, CBC, CBS News, The Time, France24, Channel News Asia, News24, Global News, Deutsche Welle, UN News, Politico, Eureporter, Euobserver, Breitbart, Huffington Post, Sky News, Realitatea, Știripesurse, TVR, Hotnews, Romania Liberă. Among the consulted agencies some are assumed to be ideologically oriented to the left or right of the political spectrum, however, we chose to keep them for the exercise of researching the proposed solution and for the confrontation of ideas on the same topic. Some problems encountered were: alteration of the title and/or content with the institution name, which can distort the clustering (example Thomson Reuters source), the use of click-bait words, so we introduced an additional processing step of the news headlines before inserting them into the database, the presence of entertainment or commercial news that do not cover any person, organisation, or event of national or world importance, and the lack of free access to some news sources that request a paid subscription (example The Wall Street Journal).

Some external *APIs* were needed, like *DeepL Translator*⁶. This translation API has been integrated to translate articles from Romanian to English. The translation

⁶ <https://www.deepl.com/translator>

from Romanian is necessary for the analysis of the news, the classification, and the extraction of the entities. All these algorithms use lexicons of the English language or neural networks trained on English texts. If we use untranslated text, then the results are poor. The functions that compile the list of entities for each item do not work well on untranslated text. This collection can facilitate further application development through multilingual support based on the use of the *DeepL* translation service, increasing accessibility to a wider audience.

4 Clustering and Summarisation

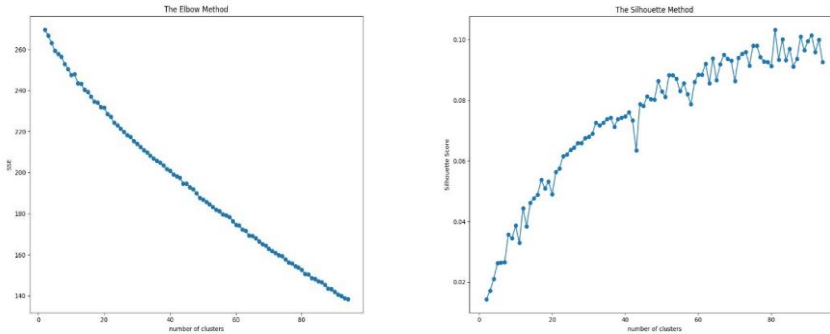
Currently, the most used news sources are news sites, *social media*, and *video streaming* platforms, which have outranked more traditional ways of information. That is why we need a joint effort from the search engines, *social media* applications, as well as media institutions active in the online space, to promote fair and honest journalism and to reduce visibility and flag sources of false information that do not respect journalistic principles. It is not easy to determine whether an article is fake, true, or satirical (Bachmann *et al.*, 2022).

4.1 Clustering

Our approach to this problem is placing the main press articles into thematic groups, summarising them, and extracting entities and their semantic analysis, through which the reader can compare the various sources of information and sense whether the text is written with too much subjectivity or bias. Thus, the average user can conclude that a news article is false, biased, or at least of poor quality, either if the information transmitted is partially or unverifiable because it does not match the accounts of other observers, or if the author intends to mislead the reader by using emotional, imprecise, strongly negative or positive expressions. Regarding the field by which the news is grouped, we have experimented with several variants. We have tested clustering by analysing the full text of the news, but the downside is it increases the computing time of clusters. We tested clustering by the summary of the articles which led to a good result as a computational cost-accuracy ratio. A small percentage (less than 5%) of articles remain not well clustered, but usually, they are filtered out because the topic (cluster) associated is marginal (has few elements). We consider the cause to be the recurrence of some common words in the content of the sentences. They raise the similarity score between certain articles, placing them in the same group, even if they have a different topic. A positive aspect of clustering regardless of the field we cluster by is that it is not influenced by language. The algorithm is as effective on the Romanian text as on the English one because the grouping is done

by calculating similarities between words, and their frequency, without taking into account the language in which they are written.

An increase in the accuracy of the grouping was obtained by focusing on the field of entities, previously extracted from English language news. By resorting to them we eliminate the problem of common words in the language because we only check the similarity between the names of people, places, institutions, and others that appear in the news. The grouping is done using the K-means algorithm, so we first look for the optimal number of newsgroups. We calculate this number by iterating progressively through several possible values, for example between a third and a half of the number of items. At each iteration, we calculate the sum of squared errors (SSE) and observe how it decreases with the increase in the number of groups. We track when this decrease becomes insignificant and then we choose the optimal number of clusters. Another variation is to calculate at each iteration, for each item, how close it is to its own group compared to the other groups. This method, called silhouette score, tracks the average score for each possible number of clusters and chooses the one that gives the maximum score (see Figure 3).



(a) Using the elbow method

(b) Using the silhouette method

Figure 3. Finding optimum clusters number.

The last field we tested for clustering is the description. It is best suited for Romanian articles, as we can skip their translation and entity extraction. When the description is not available, the title takes its place, and the results are satisfactory, with a below 10% error rate.

4.2 News Summary

There are two methods of summarising a text. The first of them, the extractive method, seeks to identify the most important sentences in the text through algorithms such as Text Rank and render them, leaving the rest. We have tested

several models available in the python transformers collection that are trained for the task of extractive summarisation. For example, Bert model gives a fair abstract when given an article, but usually misses out on important information, like names of people or places involved in the event. We have tested an algorithmic approach, extracting the entities from the text, and then the sentences where they occur. The result is more precise than the extractive summary, but it tends to be too long. Most of the sentences in a press article contain an entity. In the end, we settled for a superior model: T5 (Text-To-Text Transfer Transformer), made available by Google, which could be trained for the task of abstractive summarisation. Our goal was to train the model on a data set in order to make it capable of generating a summary that conveys the main ideas of the original text as if it were written by a human. While extractive summarisation is generally faster and requires less computation, abstractive summarisation can produce more fluent and coherent summaries but is computationally expensive. The T5 model was built for a wide range of tasks like classification, translation, summarisation, text filling, and answering questions (Roberts *et al.*, 2020). It comes in several versions, and we used the t5-base one. The model is pre-trained on the C4: Colossal Clean Crawled Corpus, a huge content of over 750 GB of written English text. But in order to increase its capabilities in a certain task you have to train it further. That's why we took the *CNN/DailyMail* data set, version 3.0.0, having over 300.000 news articles and their summary, and trained the T5 model for the summarisation task.

Some limitations for the data set are the language used - English, constraining us in development to use the translation API and the length - most of the articles are short, around 1000 chars, thus generating in development the problem of repetitions when summarising longer articles (See *et al.*, 2017). The training was conducted in the *Google Colab Pro* cloud, due to its high computational cost, and it made use of about a quarter of the big data set. Some of the improvements made to the model can be seen in Table 1.

Table 1. Training results.

Metrics	<i>T5-base</i>	<i>T5-base trained</i>
evaluation loss	14.32997798	0.23593537
evaluation run-time	85941.01	64612.55
evaluation samples per second	0.156	0.207
evaluation steps per second	0.019	0.026

After extensive training, our T5 was able to generate a summary with increased efficiency and accuracy than the pre-trained model made available by Google. We observe a big decrease in the evaluation loss and the run-time, which indicates that the model used has learned to summarise news efficiently and accurately.

Beyond numbers, we have tested numerous articles and examined how they are resumed. One problem identified is that longer articles tended to generate repetitions in the abstract. To overcome this, we split longer than 14 sentence articles into two or more parts, summarised them individually, and finally concatenated the results. The results are remarkable: the summaries are concise and precise, getting close to the summaries given by state-of-the-art models like *OpenAI ChatGPT 3.5*. Comparing the options available: Bert, T5-base, T5 trained on the *CNN/DailyMail* data set and T5 trained with the splitting of longer texts, the last one we considered the best.

4.3 Testing and Evaluation

The News in Brief went through various stages of development through User Testing: 15 people, both with and without technical training, used the application and directly provided suggestions. Most were in the area of the graphical interface which led to its considerable improvement during development. Another part of the feedback was related to the functionality of the application and determined: (1) increasing the period of time from which we collect the news from two days to one week; (2) which led to the considerable growth of the database to over 3000 press articles; (3) adding the search function; (4) the diversification of the attributes by which the search is made - it started from the title, description, and summary and added the keywords, thematic classification, web domain; (5) changing the criteria by which the newsgroups are sorted, finally they are ordered by date and by the number of news from unique sources; (6) improving the loading time of some views by displaying available data early and processing expensive calls asynchronously.

Currently we are pondering the copyright aspects involved and sharing insights with fellow journalists in order to assess the best way to use our work: either as a technical research tool, or as a general use app. For sure, users participating in testing appreciated the extraordinary ability to synthesise so much information and points of view in a simple, engaging way and the short time in which you can be updated about the world in which we live.

5 Conclusions

With the help of the proposed solution, the user can weigh the various angles from which an event is viewed and reported. In this way, we believe that every observer has the right to have a voice and to dispute the mainstream narrative in an argued and honest manner. Automated methods of grouping and summarising are not perfect, but they are powerful tools both for ordinary people and for specialists who follow the evolution of mass media and journalists who expose the truth to

the public. These methods can be improved both individually and as a combination scheme. A good improvement would be to use some of the state-of-the-art language models with multilingual and cross-lingual methods so that translating text, which leads to information loss, would not be necessary. Also, new press agencies may be added, possibly a local news registry depending on the user's geographic location. Integration with the APIs of some social networks could be developed, so that in addition to the news about a certain topic, in a cluster we could also have some posts from the personalities involved in that event.

References

- Abonizio, H. Q., de Morais, J. I., Tavares, G. M., and Barbon Junior, S. 2020. Language-Independent Fake News Detection: English, Portuguese, and Spanish Mutual Features. *Future Internet*, 12(5).
- Atodiresei, C.-S., Tănăsescu, A., and Iftene, A. 2018. Identifying Fake News and Fake Users on Twitter. *Procedia Computer Science*, 126:451-461. *Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 22nd International Conference, KES- 2018*, Belgrade, Serbia.
- Bachmann, P., Eisenegger, M., and Ingenhoff, D. 2022. Defining and Measuring News Media Quality: Comparing the Content Perspective and the Audience Perspective. *The Int. Journal of Press/Politics*, 27(1):9-37.
- Bawden, D. and Robinson, L. 2009. The Dark Side of Information: Overload, Anxiety and Other Paradoxes and Pathologies. *Journal of Information Science*, 35(2):180-191.
- Campbell, S. 2023. How Many News Articles are Published Every Day? *The Small Business Blog*, 4. <https://thesmallbusinessblog.net/news-articles-published-every-day/>
- Cușmuliuc, C.-G., Sava, I., Crainic, I.-D., Coca, L.-G., and Iftene, A. 2020. Automatic Fake News Identification System. *Proceedings of the 15th International Conference on Linguistic Resources and Tools for Processing the Natural Language*, Bucharest, Romania, pages 161-172.
- Cușmuliuc, C.-G., Amarandei, M., Pelin, I., Cociorva, V., and Iftene, A. 2021a. Uaics at Checkthat! 2021: Fake News Detection. *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum*, Bucharest, Romania, September 21-24, 2936:494-507.
- Cușmuliuc, C.-G., Susan, S.-C., Chirica, B.-D., and Iftene, A. 2021b. News Identification Metric for Classification Prefiltering. *International Conference on INnovations in Intelligent SysTems and Applications (IEEE INISTA 2021)*, pages 1-5.
- Gifu, D. and Cristea, D. 2012. Multi-dimensional Analysis of Political Language. *Lecture Notes in Electrical Engineering*, Vol. 164(1), James J. (Jong Hyuk) Park, Victor C.M. Leung, Cho-Li Wang, Taeshik Shon (eds.), pages 213-221.
- Gifu, D. and Cioca, M. 2013. Online Civic Identity. Extraction of Features. *Procedia - Social and Behavioral Sciences*, Vol. 76, University of Pitești Publishing House 2013, Elsevier, pages 366-371. DOI: 10.1016/j.sbspro.2013.04.129.

- Lehrman, S. 2023. Trust Matters. In a Confusing World Flooded with Misinformation, Principled Journalism Must Step Forward and Earn the Public Trust. *The Trust Project*. <https://thetrustproject.org/about/> Accessed on: 10.06.2023.
- Machete, P. and Turpin, M. 2020. The Use of Critical Thinking to Identify Fake News: A Systematic Literature Review. *Lecture Notes in Computer Science* (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12067 LNCS:235–246.
- Meinert, J., Mirbabaie, M., Dungs, S., and Aker, A. 2018. Is it Really Fake? - Towards an Understanding of Fake News in Social Media Communication. In: Meiselwitz, G. (eds.) *Social Computing and Social Media. User Experience and Behavior*. SCSM 2018. LNCS vol 10913. Springer, Cham. https://doi.org/10.1007/978-3-319-91521-0_35.
- Minea, I. and Iftene, A. 2017. News on Twitter. *Proceedings CMSM4*, pages 535-538.
- Naisbitt, J. 1982. Megatrends: Ten New Directions Transforming Our Lives. Warner Books, New York.
- Nenkova, A. and McKeown, K. 2011. Automatic Summarization. *Foundations and Trends® in Information Retrieval*, 5(2-3):103–233.
- Pariser, E. 2011. The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin Books.
- Roberts, A., Raffel, C., Lee, K., Matena, M., Shazeer, N., Liu, P. J., Narang, S., Li, W., and Zhou, Y. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21.
- See, A., Liu, P. J., and Manning, C. D. 2017. Get to the point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 1073-1083, Vancouver, Canada, July. ACL.
- Steinbach, M. S., Karypis, G., and Kumar, V. 2000. A Comparison of Document Clustering Techniques. In *KDD Workshop on Text Mining*, Vol. 400.
- Valenzuela, S. and Arriagadac, A. 2014. Facebook, Twitter, and Youth Engagement: A Quasi-Experimental Study of Social Media Use and Protest Behavior Using Propensity Score Matching. *International Journal of Communication*, 8:2046–2070.
- Vosoughi, S., Roy, D., and Aral, S. 2018. The Spread of True and False News Online. *Science*, 359(6380):1146–1151. Accessed: 8.06.2023.

MYANMAR SIGN LANGUAGE RECOGNITION USING DEEP LEARNING

BHONE MYINT SWE¹, MIN THAW PHYO¹, MIN CHAN LUN¹,
AUNG SANN THIT¹, AUNG HLAING MOE¹,
KAUNG KHANT MIN¹, PHYO THU HTET², AYE SU PHYO²

¹*Yangon, Myanmar, Khaw Thwar Mel*
khawthwarmel@gmail.com

²*Mandalay, Myanmar, Simbolo*

Abstract

In this paper, we propose a word and expression-level recognition system using Pose Estimation and Long Short-Term Memory (LSTM) for Myanmar Sign Language. The model converts Myanmar Sign Language from videos into Burmese words and expressions, using diverse data collected from different people in various situations for training. In the data preprocessing stage, the dataset was subjected to the landmark of face, body, and hand using *mediapipe*, with the resulting data stored as *numpy* arrays in CSV format. These arrays were subsequently used as the training data input for the model. The model's function involved predicting or recognizing the signs presented in the input video. The accuracy rate during the training phase was 94.4%, whereas the accuracy rate during the validation phase was slightly lower at 92.78%. A 60% accuracy rate was attained in the real-time testing phase using signer data that had not been seen before by the model. Whereas people who are new to studying sign language achieve an accuracy rate of 80% resulting in human performance remaining superior to the model. Our main purpose is to break the barrier and connect hearing-impaired people to hearing people with our research and technologies as much as possible. Lastly, we believe that our approach will help solve the problem of hearing-impaired people unable to communicate with hearing people.

Keywords: Myanmar Sign Language, Deep Learning, long short-term memory (LSTM), Dynamic and Static Signs, Mediapipe, Landmarks.

1 Introduction

The project was started in response to the common communication issues between the hearing-impaired community and individuals who are unfamiliar with sign language by Khaw Thwar Mel. Khaw Thwar Mel is a social enterprise with the

aim of empowering the disabled community through innovative technological solutions, with services including Myanmar sign language learning, tailor-made technological courses, and job opportunities. The team understood the need for a solution that might let these two groups communicate effectively. The researchers set out to create a system that is able to enhance communication and interaction between people who are skilled in Myanmar Sign Language and those who are not as they believed that technology had the capacity to close this gap. The main goal is to establish a welcoming environment where everyone can communicate effectively.

There have been many languages invented and lost since the beginning of civilization ILS, International Languages Services, (Blog article, 2019). There are three forms of communication which are verbal, non-verbal, and visual Academy (TMC Academy, 2023). Most people use verbal or oral communication methods. There are more than 100 languages spoken in Myanmar and sign language is the least popular (Ko and Mikami, 2005). People who suffer from the disability of deafness and muteness (hearing-impaired people) use sign language which is a kind of non-verbal communication.

In Sign Language, people speak through actions such as facial expressions and hand gestures. Hand gestures have three crucial components which are sign, position, and movement (Napoli and Ferrara, 2021). Unfortunately, most people do not understand it. That is why it is challenging to communicate with deaf people. However, technologies are becoming more and more advanced and Artificial Intelligence (AI) has become one of the solutions that can solve communication problems.

In Myanmar, two Myanmar Sign Languages (MSL) are mainly used. One is used in Mary Chapman School for the Deaf, and another is used in the school for the deaf in Mandalay and Tamwe, Yangon (Aung *et al.*, 2020)

Sign language is divided into two categories, static and dynamic words and expressions. Static words refer to signs that involve no movement, only a particular hand shape held in a certain position. Examples of static signs include the signs for letters of the alphabet (e.g. က, ခ, ဂ like a, b, c), numbers (e.g., ၁, ၂, ၃ like 1, 2, 3), and some basic nouns (e.g., အင်္ဂါ (Tuesday), တနင်္လာ (Monday)) that do not involve significant movement. Dynamic words and expressions involve movement as part of the sign. The hands may change shape, move through space, or rotate as part of the sign. Many verbs, adjectives, adverbs, and more complex concepts in sign language are expressed using dynamic signs. Examples include word and expression like မင်္ဂလာပါ (Hello), and နေကောင်းလား (How are

you?). Video data is used for training, validation, and testing since both static and dynamic signs are expected to work with the sign language recognition system.

In this paper, we use a real-time Myanmar Sign Language recognition system based on the most-used sign language in Yangon. We used the sign language used in Yangon because of the availability of experienced signers and it is one of the popular sign languages in Yangon. Deaf people communicate with each other using sign language. But most people do not understand sign language.

There is not much research about the recognition system for Myanmar Sign Language. Most research is about recognizing static sign language. As for the training model, pose estimation and LSTM are used together since the image does not need to have a clear background and they work together well in real-life situations. In the next section, some works of others are compared and explained with ours.

The rest of the paper is as follows. Section 2 is about the related work that was used to compare and reference the paper. In Section 3, the focus is on the dataset, encompassing data collection, data comprehension, and an examination of the characteristics of the selected 10 signs. Following on, section 4 describes the system architecture used in the paper including preprocessing the data and the architecture of the model. Section 5 discusses the experimental results of a study on Myanmar Sign Language recognition using Pose Estimation and Long Short-Term Memory (LSTM). Finally, the text provides concluding statements and outlines potential directions for future research in Section 6.

2 Related Work

To begin with, the sign language recognition system (SLR) is explored by most AI research teams as AI deep learning systems become more efficient and reliable. The static sign language recognition system from (Tolentino *et al.*, 2019) uses the Convolution Neural Network (CNN) theory. The overall average accuracy of the system is 93.667% consisting of 90.04% average accuracy of letters, 97.52% average accuracy of static words including 35 words, and 93.44% average accuracy of numbers (Tolentino *et al.*, 2019). However, this approach works only for static sign language.

In another paper, Ni Htwe Aung (Aung *et al.*, 2020) developed Myanmar Sign Language Recognition for Myanmar Consonants using Transfer Learning. They used up to 12 pre-trained models, using the TensorFlow library. Among 12 models, VGG16, ResNet50, and MobileNet with epoch 50 resulted in the highest accuracy score at 94%. However, their Myanmar Fingerspelling Recognition system can only recognize 31 static but lacks 2 dynamic sign language characters (Aung *et al.*, 2020).

Following on, Shivanarayna Dhulipala (Dhulipala *et al.*, 2022) wrote a paper about Sign and Action Detection using Deep Learning. Their study aims to develop a qualified deep-learning model for British sign language. In their research, they created two different models with CNN and LSTM. Between those two, the CNN model yields the highest performance which results in training accuracy at 98.8% and validation accuracy at 97.4%. In contrast, the performance of the LSTM model is really poor, and the model only achieved respective accuracies at 49.4% for training and 48.7% for validation at most.

Roman Töngi (2021) refers to using transfer learning to recognize sign language. He based his model on the Inception-v3, 2D ConvNet, and ImageNet pre-trained models. The American Sign Language dataset MS-ASL was used as a source dataset to pre-train a model, then it was fine-tuned separately on the German Sign Language dataset SIGNUM. By using the transfer learning method, the highest accuracy achieved is 75%. However, they did not try to apply his models to real-time testing.

Lastly, in (Kothadiya *et al.*, 2022), the authors proposed a deep learning-based model, utilizing a combination of LSTM and GRU, which demonstrates an impressive accuracy rate of approximately 97% in recognizing Indian Sign Language (ISL) gestures, as evidenced by the experimental results. This model's strengths lie in its ability to effectively detect and interpret sign language in real-time scenarios, offering a valuable communication tool for individuals with speech or hearing impairments. However, like any computational model, it has its limitations, including the need for a substantial dataset for optimal performance and potential challenges in dealing with variations in lighting conditions and angular input frames. Further research and data refinement are required to address these cons and enhance the model's robustness for practical applications in diverse environments.

Long Short-Term Memory (LSTM) is widely used in various applications that involve sequential data. For instance, in (Drewil and Al-Bahadili, 2022) we have a deep learning module of LSTM (long short-term memory) and Genetic Algorithm (GA) to find the best parameter and pollution level for the next day. In our study, we employ LSTM to identify and interpret sign language gestures which include dynamic words, and expressions.

3 Dataset

3.1 Data Collection and Data Understanding

Meanwhile, there are no pre-recorded datasets for Myanmar Sign Language that are publicly available, so we manually collected the datasets. The data was collected for 10 signs. The details of the 10 words are explained in section 3.1.1.

The collected data include 3 deaf participants who were recruited from different backgrounds, ethnicities, and religions. All participants are fluent in Myanmar Sign Language (MSL) and have been deaf since birth or early childhood.

When the data was collected, the participants were recorded in a controlled setting to get better data quality. The green screen was used as a background so that each individual can be seen clearly in the video. Two lighting sources were set up and the same lighting setting was used from the start till the end so that there are no extra shadows in the data. We also made sure the external light sources were not interfering with the setting. Then, we set the setting with Sony A7siii with lens 24-70 GM. The camera settings were set constant, which included shutter speed 1/50, F2.8, exposure +1, and ISO 640. During the data collection, The ISO was changed between 640 and 800 to get the right exposure since the exposure was changed according to the situation. The camera position was slightly adjusted up and down in order to have each participant inside the same video frame.

Generalized datasets are collected to make sure that there are no problems in training the AI. As Sign Language only needs the upper part of the body, only the required area was recorded. The data was collected in different clothes and conditions but in the same setting environment. The participants changed their clothes with different colors 3 times. Their facial expressions and hand placements were made sure to be consistent. Each word includes 90 videos and has 900 videos for 10 signs. There are 10 signs used in the paper and 90 videos are included for each word so that it has 900 videos in total.

Nature of the selected words and expressions

Subsequently, ten dynamic signs with a variety of hand movements and facial expressions were chosen for the data collection. The chosen words and expressions are shown in Table 1. In sign language, hand signs, positions, and facial expressions are crucial because a change in one sign or facial expression causes another meaning. However, some words do not change their meanings by changing facial expressions. The chosen words and expressions are explained in terms of their nature as follows.

The words “Hello (မင်္ဂလာပါ)”, “Monday (တနင်္လာ)” in Figure 1, and “Tuesday (အင်္ဂါ)”, “Sunday (တနင်္ဂနွေ)” have same hand sign but different hand position. On the other hand, the words “Wednesday”, “Thursday” and “Friday”, “Saturday” in Figure 2 have the same hand position but different hand signs. Moreover, both “How are you?” and “It’s okay” have two hand movements meaning they each have two different hand signs where the initial signs are different but the final hand signs are the same but have different hand positions in Figure 3.

As for facial expression, some signs are important to have the expression while others are not. The words from no. 4 to 10 in Table 1 do not need to have facial expressions and it is not crucial. But the words and expressions from no. 1 to 3 are important to have facial expressions since without the facial expression or wrong facial expressions defines other meanings. The word “How are you?” is a question that has facial expressions but if it is not included it means “I’m fine.” or something else.

Moreover, the word “It’s okay.” is also important to have facial expressions. Without the expressions or changing the expressions into the same expression as “How are you?” turns it into other meanings like “Are you okay?” or something else. Overall, understanding the nature of hand movements and facial expressions in sign language is crucial for accurately conveying meaning, and the ten chosen signs provide excellent examples for training AI models.

The dataset gathered for the research paper will be made publicly available on Kaggle¹, and provided in a CSV file format for Myanmar Sign Language Recognition. Given the scarcity of such datasets in Myanmar, this release is expected to facilitate and support forthcoming advancements in this field.

Table 1. Words and expressions used for the data collection.

Ebola. No	Words
1.	မင်္ဂလာပါ (Hello)
2.	နေကောင်းလား (How are you?) (Expression)
3.	အဆင်ပြေတယ် (It’s okay.) (Expression)
4.	တနင်္လာ (Monday)
5.	အင်္ဂါ (Tuesday)
6.	ဗုဒ္ဓဟူး (Wednesday)
7.	ကြာသပတေး (Thursday)
8.	သောကြာ (Friday)
9.	စနေ (Saturday)
10.	တနင်္ဂနွေ (Sunday)

¹ The largest AI ML community available at <https://www.kaggle.com/>

3.2 Comparison for Selected Words and Expressions



(a) Hello



(b) Monday

Figure 1. (a) and (b) Same hand sign different hand position.

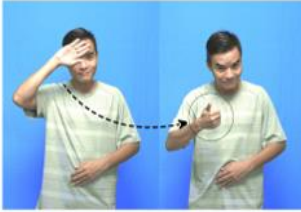


(a) Friday



(b) Saturday

Figure 2. (a) and (b) Different hand signs same hand position.



(a) How are you?



(b) It's ok.

Figure 3. (a) and (b) Same Final Hand Sign.

4 System Architecture

4.1 Data Preprocessing

Subsequent to the data collection phase, various editing techniques were employed to enhance the overall quality and accessibility of the dataset. These editing methodologies encompassed a spectrum of modifications, including but not limited to adjustments in the background color, manipulation of brightness and contrast, fine-tuning of exposure levels, and the resizing and flipping of the original 4k resolution footage to 1080p, 720p, and 480p, respectively as in Figure

4. This multi-faceted approach was implemented with the intention of broadening the applicability of the data, ensuring seamless compatibility across diverse electronic devices and platforms.

The collected video data is used to detect landmarks which are later stored as *numpy* arrays using *mediapipe* in CSV file format. The landmarks refer to key points or positions on a detected part, such as facial landmarks representing specific facial features. *Mediapipe* extracts landmarks from the body, face, and hands for sign language recognition by utilizing dedicated deep learning models for each body part as shown in Figure 5. The extracted landmarks are really important because they help the system understand sign language better. They let the system see how the signer moves, makes gestures and expresses themselves. This special data is then used to teach the system what different signs mean in sign language.

Moreover, the transformative process was extended beyond mere technical alterations. An inclusive consideration was incorporated into the editing procedure, recognizing the significance of accommodating both right-handed and left-handed individuals within the hearing-impaired community. Consequently, a deliberate mirroring or flipping of the visual data was executed to ensure that the AI algorithms were proficient in interpreting and comprehending the subtleties of both right and left-handed sign language expressions, thereby fostering an environment of inclusivity and accessibility for all users. Since each video is 5 seconds long, there are 150 frames in each. Due to the limitation of computation power, RAM, and GPU, 3 frames are stored as 1 frame using mean calculation resulting in 50 frames in CSV file format for each video.

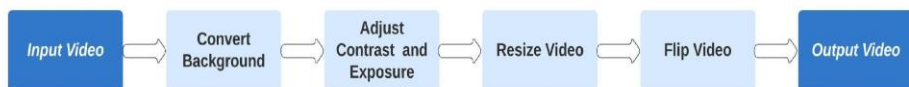


Figure 4. Image preprocessing steps.

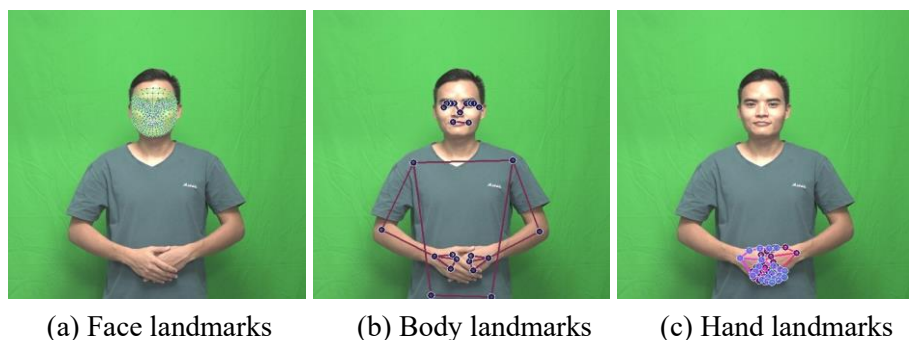


Figure 5. Landmarks.

5 Model Architecture

The model architecture uses sequential layers, the container for the model layers which also defines a linear stack of layers where you can add one layer at a time. Three LSTM (Long Short-Term Memory) layers are used in this model. LSTM is a type of recurrent neural network (RNN) designed to handle sequences of data. The LSTM model is used because of the video data. Three fully connected (Dense) layers are used after the LSTM layers.

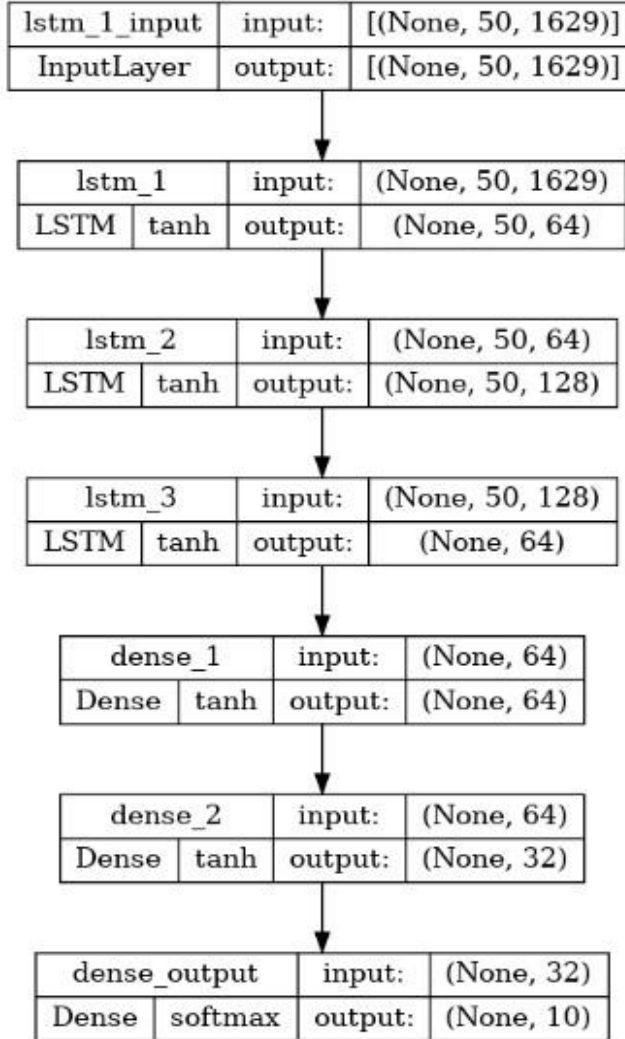


Figure 6. Architecture.

As shown in Figure 6, the ‘Sequential()’ function creates a sequential model for layer addition in sequence. ‘LSTM()’ adds LSTM layers, where the first layer (‘lstm_1’) has 64 units, returns sequences, and uses the hyperbolic tangent (tanh) activation function, expecting an input shape of (50, 1629) representing 50 frames and 1629 landmarks. The second layer (‘lstm_2’) has 128 units, also returning sequences, with a *tanh* (tangent hyperbolic) activation function. The third layer (‘lstm_3’) has 64 units, doesn’t return sequences (only the final hidden state), and utilizes the tanh activation function. ‘Dense()’ adds fully connected dense layers, with ‘dense_1’ having 64 units, ‘dense_2’ having 32 units, and ‘dense_output’ having the same number of units as the ‘words’ variable (representing the number of target classes). The activation function for ‘dense_1’, ‘dense_2’, and ‘dense_output’ is tanh, except for ‘dense_output,’ which uses softmax activation for multiclass classification.

6 Experimental Results

The total dataset consists of 900 videos demonstrated by 3 signers. Training and validation data are then separated by 80% and 20% respectively. Videos of a signer who is not included in those data are used for later testing. The accuracies of training and testing results are shown in Table 2. The accuracy score macro is 92.78%. Precision score macro is 93.16% and recall score macro is 92.78%. Lastly, the f1-score macro is 92.85%. The accuracy of the training dataset is 94.44%. The validation accuracy is 92.78%. Due to our limitations, only 3 expert signers are used for the datasets and it is not generalized. If other expert signers demonstrate exact signs, the model can predict well but inaccurate signs from the deaf can lead to wrong predictions.

In Figure 7, the confusion matrix serves as a visual aid in machine learning, depicting how well the classification algorithm performs. It offers a detailed overview of predicted versus actual classifications, including true positives, true negatives, false positives, and false negatives. Analysing the matrix provides valuable insights into the accuracy and efficacy of the model’s predictions. Notably, among the words and expressions evaluated, “Hello,” “How are you?”, and “It’s okay” demonstrate the highest accuracy with a count of 18, whereas the word “Saturday” yields the lowest result.

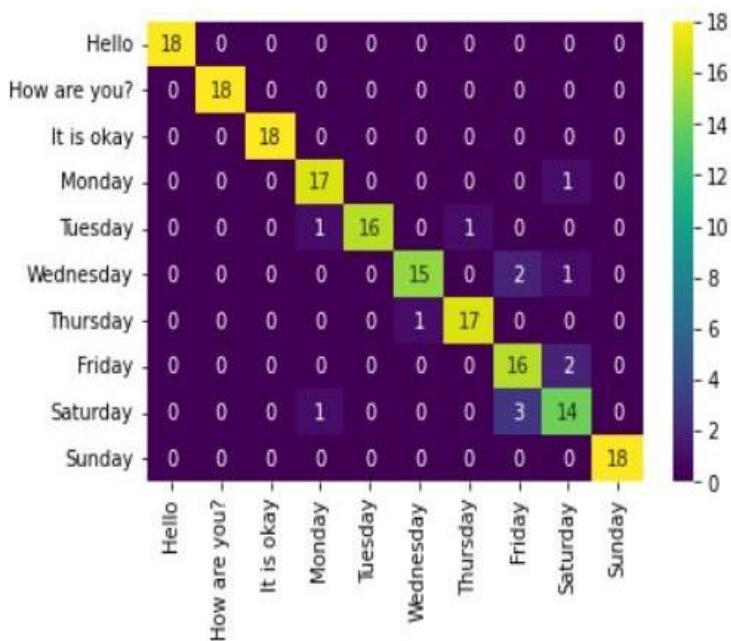


Figure 7. Confusion matrix.

Table 2. Accuracy of training datasets using 80% of total datasets.

No	Words and Expressions	Precision	Recall	F1-score	Support
1	မင်္ဂလာပါ (Hello)	1.00	1.00	1.00	18
2	နေကောင်းလား (How are you?) (Expression)	1.00	1.00	1.00	18
3	အဆင်ပြေတယ် (It's okay.) (Expression)	1.00	1.00	1.00	18
4	တနင်္လာ (Monday)	0.89	0.94	0.92	18
5	အင်္ဂါ (Tuesday)	1.00	0.89	0.94	18
6	ဗုဒ္ဓဟူး (Wednesday)	0.94	0.83	0.88	18
7	ကြာသပတေး (Thursday)	0.94	0.94	0.94	18
8	သောကြာ (Friday)	0.76	0.89	0.82	18
9	စနေ (Saturday)	0.78	0.78	0.78	18
10	တနင်္ဂနွေ (Sunday)	1.00	1.00	1.00	18

In Figure 8, the model wrongly predicts the gestures for “Friday” as “Saturday”. The hand gestures for “Friday” and “Saturday” appear quite alike, sharing similar hand positions, facial expressions, and movements. The key distinguishing factor lies in the number of fingers placed on the cheek, where “Friday” displays three fingers, while “Saturday” is only one. Occasionally, when users provide particularly ambiguous instructions, the system’s predictions are not consistently accurate. These observations underscore the importance of enhancing the system’s ability to discern subtle differences and adapt to various scenarios, thereby enhancing its overall performance and reliability. The signer’s data is not part of the training or testing datasets in the paper, meaning it’s new to the model.



Figure 8. Wrong prediction.

In Figure 9, the model predicts the expression “It is okay” correctly. The expression “It is okay” and the word “Wednesday” exhibit distinct hand positions and gestures that do not coincide with those of other words, demonstrating accurate prediction. This suggests the model’s efficacy in distinguishing between different signs and accurately translating them into their corresponding words and expressions.

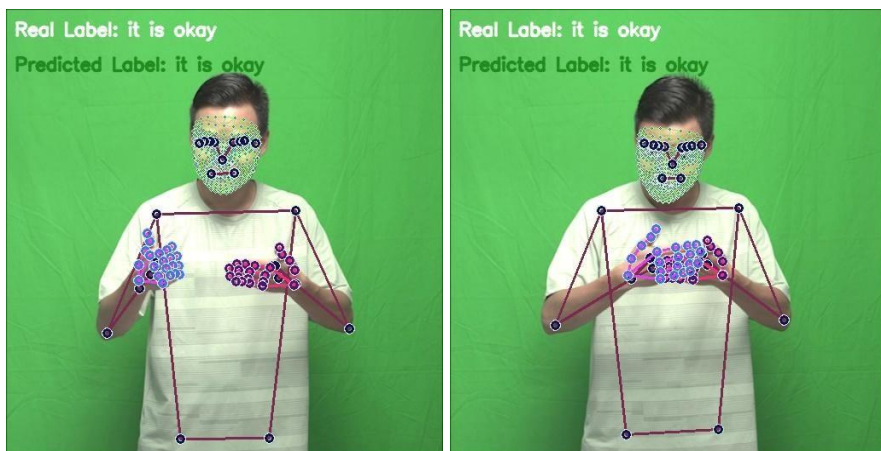


Figure 9. Correct prediction.

7 Conclusion

In this paper, the data are collected from 3 deaf signers for the word-level recognition system for Myanmar Sign Language (MSL) using Pose Estimation and Long Short-Term Memory (LSTM), aiming to bridge the communication gap between hearing-impaired individuals and the general population. Through the manual collection of data encompassing both static and dynamic signs, and the utilization of advanced AI technologies for data preprocessing and model development, the system achieves a promising accuracy of 92.78% in recognizing MSL words and expressions. While this research represents a significant step towards facilitating communication for the hearing-impaired, challenges such as limited signer diversity may impact prediction accuracy. Nevertheless, this work lays a foundation for future advancements in MSL recognition and the development of technology-driven solutions to enhance the lives of hearing-impaired individuals.

Acknowledgements

We would like to thank the ten volunteers from LLDD for giving not only their precious time but also their studios for the data collection. We would also like to express our gratitude to Thonenari and their camera crews. We are very appreciative of Literacy and Language Development for the Deaf (LLDD) and Mr. Soung Yein permission to use his photos in the paper and helping us collect the related data. Finally, our sincere thanks go out to all the participants involved throughout the entirety of this research, as well as to Simbolo and Khaw Thwar Mel.

References

- TMC Academy. 2023. Business Communication and ICT. Singapore: British University College.
- Aung, N. H., Thu, Y. K., Maung, S. S., Moe, S. Z., and New, H. M. 2020. *Transfer Learning Based Myanmar Sign Language Recognition for Myanmar Consonants*. PhD thesis. MERAL Portal.
- Dhulipala, S., Adedoyin, F. F., and Bruno, A. 2022. Sign and Human Action Detection Using Deep Learning. *Journal of Imaging* 8.7 (2022), p. 192.
- Drewil, G. I., and Al-Bahadili, R. J. 2022. Air Pollution Prediction Using LSTM Deep Learning and Metaheuristics Algorithms. *Measurement: Sensors* 24, December 2022, 100546.
- International Languages Services ILS. 2019. *Understanding Extinct Languages: When and Why They Die Off*. Blog article, November 12th, 2019. <https://www.ilstranslations.com/blog/understanding-extinct-languages-when-and-why-they-die-off/>
- Ko Ko, W. and Mikami, Y. 2005. Languages of Myanmar in Cyberspace. In *Proceedings of TALN & RECITAL 2005*, pages 269–278.
- Kothadiya, D., Bhatt, C., Sapariya, K., Patel, K., Gil-González, A.-B., and Corchado, J. M. 2022. Deepsign: Sign Language Detection and Recognition Using Deep Learning. *Electronics* 2022, 11(11), 1780.
- Napoli, D. J. and Ferrara, C. 2021. Correlations between Handshape and Movement in Sign Languagesp. *Cognitive Science* 45.5, 2021, e12944.
- Tolentino, L. K. S., Serfa Juan, R. O., Thio-ac, A. C., Pamahoy, M. A. B., Forteza, J. R. R., and Garcia, X. J. O. 2019. Static Sign Language Recognition Using Deep Learning. *International Journal of Machine Learning and Computing*, 9(6):821-827.
- Töngi R. 2021. Application of Transfer Learning to Sign Language Recognition Using an Inflated 3D Deep Convolutional Neural Network. <https://arxiv.org/abs/2103.05111>

A ROMANIAN BERT MODEL FOR LINGUISTIC ANALYSIS

RADU ION

*“Mihai Drăgănescu” Romanian Academy Institute for AI
13 “Calea 13 Septembrie”, Bucharest 050711, Romania
radu@racai.ro*

Abstract

This paper documents the creation and evaluation of a Romanian BERT model that is suitable for linguistic analysis of its encodings. To be useful from this point of view, we have built it on top of a newly developed WordPiece tokenizer that properly splits Romanian words, taking care of clitic splitting, diacritic normalisation and automatic enforcement of the Romanian Academy writing norms concerning vowel “ă” and some forms of the verb “to be”. The BERT model is currently training with the Masked Language Modeling objective on the CoRoLa corpus (approx. 760 million words), and it is going to be comparatively evaluated by studying how it behaves when predicting words and, how it encodes the plural number of selected nouns.

1 Introduction

Natural Language Processing is commonly done nowadays using the Transformers (Devlin *et al.*, 2019) architecture, which quickly obsoleted most previous state-of-the-art approaches for Question Answering, Machine Translation, Dialog Systems, Text Summarisation, and so on, the list is long. The Large Language Models (LLM), as these models are called are extraordinarily good at quickly adapting, when further trained on relatively few data (compared to the data they have been trained on), to any task related to language processing and understanding. GPT-4 (OpenAI, 2023), the latest LLM from OpenAI which is rumoured to have over a trillion parameters¹ (by comparison its predecessor, GPT-3.5 has 175 billion parameters), is the backend for the now world-famous ChatGPT dialog system² which can answer any question and elaborate intelligently on any request that is presented to it (e.g., “Give me the recipe for a pepperoni pizza.” or “Explain the Schrödinger equation.”)

¹ Value obtained by asking Microsoft’s Copilot about it.

² <https://openai.com/blog/chatgpt>

While the NLP community is trying to achieve better results at various tasks using these LLMs, the open question of how the Transformers learn the structure of a language and where is the linguistic information stored into different inner layers of the network or in the output encodings has remained largely unanswered. The power of the Transformer architecture comes from the attention mechanism³ (Vaswani *et al.*, 2023), which *learns the positional co-occurrence information of any two tokens* inside a 512-token⁴ window of text.

When supplied with billions of tokens, and when the model has hundreds of billions of parameters, the combinatorial explosion of token pairs seems to not pose a problem for the network which learns linguistic phenomena such as word formation, grammatical categories, syntactic analysis, and semantic combinatorial constraints (e.g., theta roles and multi-word expressions). If all these linguistic phenomena that are present in the trained model can be somehow pinpointed and used in isolation, the Transformer model can acquire new uses for computational linguistics, language learning, and so on. Furthermore, if the inner workings of the Transformer model would be understood, the attention mechanism could be remodelled to better reflect a specific language analysis process, e.g., dependency parsing.

There are recent studies that begin to look at what and how transformers learn the language. One such study (Bingzhi *et al.*, 2023) looked at the ability of the Transformer model to recognize long-distance subject-verb agreement. The model is compared to a LSTM neural network, doing the same task and the Transformer beats the LSTM network by more than 12%. However, Bingzhi *et al.* aim to determine if the Transformer model can encode different structures for different types of agreements, and not what are the model's parameters that attend to those structures.

Jo and Myaeng (2020, Figure 3) provide accuracy heat maps of different Transformer models attention heads (from 1 to 12 or 16, depending on the model) at different layers (1 to 12 or 24) for different linguistic classification tasks such as word order, top level constituent heads, etc. They show that different attention head outputs (not necessarily the ones at the top, from which the final token encoding is extracted) are better at different linguistic classification tasks, which provides evidence of what linguistic representation is learned at different Transformer layers. What is interesting from Figure 3 of the cited paper is that different Transformer models (large and base BERT, GPT, GPT-2 and ELECTRA)

³ The attention mechanism is to text what the convolutions are to images: ways of automatically inducing new (hidden) features from the input text, in an unsupervised manner, as the convolution window would reduce the size of an input image by applying a function on small rectangles composing it.

⁴ This is the common length of a Transformer model input.

learn the same linguistic representations (relevant to the targeted classification tasks) differently, at different layers.

While there are a lot of studies that probe the ability of Transformer models to learn certain linguistic structures, none of them focuses on how the specific structure is learned, e.g., what are the layers in the models that are responsible for identifying the specific structure, and if the same structure is learned the same way (in the same layers) by different Transformer models, making the Transformer a possible candidate for a universal language learner. Thus, the question “Is the Transformer model a universal language learner?” is a very good and important question to ask.

The present paper focuses on understanding (a very small part) of the inner workings of a Transformer model, namely on identifying the dimensions of the vector representation of a word that are responsible for the encoding of the plural number in Romanian nouns. We are going to use a BERT model that is developed specifically for dealing with properly tokenized Romanian text and we will compare it with the most used, existing Romanian BERT model (Dumitrescu *et al.*, 2020). In what follows, in Section 2 we present a WordPiece tokenizer that we developed specifically for dealing with Romanian, in Section 3 we detail the Romanian BERT model we intend to use for further linguistic analysis and in Section 4 we present our quantitative analysis of vector dimension correlations of selected Romanian plural nouns. Finally, the paper ends with our conclusions in Section 5.

2 A Romanian WordPiece Tokenizer

To study how morpho-syntactic attributes are represented in Romanian, we need a tokenizer that has the following properties:

1. automatic normalisation of BERT model inputs, so that the correct diacritics are used for "ș" and "ț" ("ş" and "ţ" are the correct characters).
2. enforcing of the correct spelling in accordance with the requirements of the Romanian Academy, namely the forms of "sunt", "suntem", "sunteți" (instead of "sînt", "sîntem" and "sînteți") and writing with "â"⁵. We need this type of behaviour because the training of the BERT model is done on the CoRoLa corpus (Barbu Mititelu *et al.*, 2019) which contains texts written in the old spelling, and the vocabulary we use contains only words spelled in accordance with the current writing norms.

⁵ In accordance with the rules stated here: https://acad.ro/grafieLimbaRomana/pag_grafieLR_norme.htm.

3. usage of a (much) richer vocabulary than the BERT models for English. English is not an inflectional language and thus, a vocabulary of 30,000 – 50,000 words is sufficient. Romanian is a language with many more inflectional forms for a word (e.g., a noun with a regular inflection usually has 8 forms, while in English, a noun only has 3 forms).
4. separate clitic pronouns, prepositions, etc. from the words to which they are hyphenated. These are well-defined words in Romanian, and they must be used to train the model.

A WordPiece tokenizer (Song *et al.*, 2021) works like this:

1. calls the text normalisation routine on input and obtains the normalised version.
2. calls the text segmentation routine according to the specifics of the implemented language. For English, a whitespace tokenization is done and then, the punctuation is separated from the tokens that contain it. A routine is then applied that removes all diacritics from words containing diacritics and converts all words to lowercase.
3. for each separate token, if it is in the vocabulary, it is given the corresponding ID (each token in the vocabulary has an associated ID – a positive integer – that is unique).
4. if the token is not in the vocabulary, it is split into two sub-tokens so that both are in the vocabulary. Splitting an unknown (non-vocabulary) token into its component sub-tokens was learned by the WordPiece tokenizer from training on a corpus, studying words with less than a threshold frequency – usually 2 – by splitting them in such a way so that the component sub-tokens are very frequent in the corpus.

The Romanian WordPiece⁶ type tokenizer implements all the properties stated above (normalises the input text, imposes the Romanian Academy writing norms, and tokenizes clitic words) and uses a vocabulary suitable for Romanian: 500,000 words. We chose this vocabulary size for the following reasons:

1. constructing the frequency list of the word forms occurring in the CoRoLa corpus and sorting it in descending order by the frequency of occurrence, the first 500,000 occurring word forms have a frequency greater than or equal to 5. We considered the frequency equal to 5 as a good threshold for the WordPiece tokenizer to learn how words (affixes and inflections) are formed from those less frequent than 5. For English this value is 2, but for Romanian we considered that we need more words to be able to automatically learn the inflection of the Romanian language.

⁶ <https://github.com/racai-ai/ro-wordpiece-tokenizer>

2. studying a lexicon of the Romanian language with over 1 million inflected word forms, if we calculate how many forms a lemma has, on average, we get the value of 9 (8.682). If the English language has 2 forms/lemma (1.673, Brysbaert *et al.*, 2016), on average, we deduce empirically that the Romanian vocabulary must be at least $8.682 / 1.673 = 5.189$ times larger than the English vocabulary (50,000 tokens). As a result, the Romanian vocabulary had to contain at least 250,000 tokens, but correlating this requirement with the considerations from point 1 above, we retained the value of 500,000 words for the Romanian vocabulary.

The most frequent Romanian words of the automatically learned WordPiece tokenizer are “de”, “în” and “și” and the most frequent inflections are “##ul” (masculine, singular, nominative-accusative), “##le” (feminine, plural, nominative-accusative) and “##lor” (plural, genitive-dative). The two dashes signify that these tokens stick to the ones that precede them to form the original tokens. The Romanian tokenizer also learned morphological prefixes such as “pre” or “ne” but also multi-word expressions such as “de la” or “precum și” (which were detected in the pre-tokenization phase). The `vocab.txt` file learned from the CoRoLa corpus is on GitHub⁷.

3 A Small Romanian BERT Model

It is a standard, "small" BERT model⁸:

- accepts sequences of tokens (produced by the tokenizer described above) of a maximum length equal to 256. For the experiments we want to perform, this context size is sufficient (even for future experiments, where we analyse the semantic components of word meanings, we do not expect the semantic dependency between the words of a phrase to exceed this value in length).
- has 4 hidden Transformer blocks (L=4).
- the resulting encoding has 256 positions (H=256).
- there are 8 attention heads (A=8)

The model is trained on a cleaned version of the CoRoLa corpus which has 35,999,401 phrases and 763,531,321 words⁹. At the time of writing, training had to be restarted at 42%, after the model had trained for 535 hours on an NVIDIA Quadro RTX 8000 graphics card with 48 GB of RAM. The training process could have used a batch size of 24 samples (instead of 16), to utilise all the available memory of the video card.

⁷ <https://github.com/racai-ai/ro-wordpiece-tokenizer/blob/master/model/vocab.txt>

⁸ <https://github.com/racai-ai/ro-corola-bert-small>

⁹ Counted with the Linux `wc -w` command.

The last checkpoint saved (intermediate model, before the crash) was named “1,279,000”, which is the number of the training batch after which the save was made (for reference, training is done on 3,072,175 batches of 16 examples each). A batch contains 16 contexts of words that have been automatically replaced with the special token “[MASK]” and for which the BERT model is trained to maximise its replacement probability with the original word. This training regime is known as the “Masked Language Modeling” (MLM) objective.

The perplexity of the model on the development set (which represents 1/1000 of the total words in the corpus) is 2.18 (calculated with the checkpoint 1,279,000), decreasing from the initial value of 5.84. On the training set, the current perplexity is 2.31, down from the initial value of 5.82. Figure 1 shows the evolution of the perplexity on the development (blue graph) and training (orange graph) sets, during the training process.

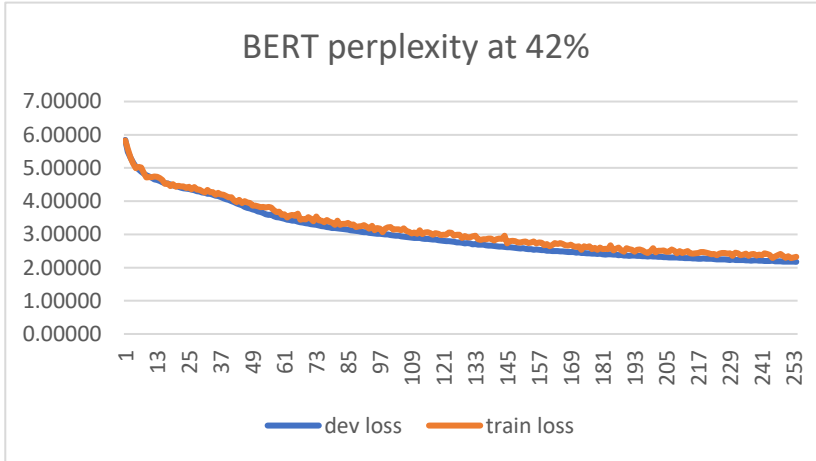


Figure 1. BERT model perplexity after training on 42% of the train set.

To evaluate how good the model is at suggesting which word fits best in a given context, we evaluate the “1,279,000” checkpoint on some test sentences. The first test phrase is “El [MASK] urcat în tren.” and for this sentence, the five most likely replacements for the token “[MASK]” are:

Our small BERT model	Bert-base-romanian-uncased-v1
El a urcat în tren.	El a urcat în tren.
El am urcat în tren.	El este urcat în tren.
El au urcat în tren.	El sa urcat în tren.
El ați urcat în tren.	El doar urcat în tren.
El e urcat în tren.	El era urcat în tren.

We see that the model correctly suggested the auxiliary verb “a” at the first position of the returned list for the Romanian tense “perfectul compus” (present perfect), in agreement with the subject “El” (He). The existing Romanian BERT model “bert-base-romanian-uncased-v1” (Dumitrescu *et al.*, 2020) presents the same suggestion as its first one, but we notice a spelling error on the “El sa urcat în tren.” where the reflexive clitic “s-” is missing the dash.

If we want to test the reflexive nature of the verb “a se urca” (to board), we try the phrase “El [MASK] a urcat în tren.” and we get:

Our small BERT model	Bert-base-romanian-uncased-v1
El s- a urcat în tren.	El nu a urcat în tren.
El l- a urcat în tren.	El tocmai a urcat în tren.
El și- a urcat în tren.	El doar a urcat în tren.
El i- a urcat în tren.	El abia a urcat în tren.
El mi- a urcat în tren.	El chiar a urcat în tren.

The reflexive clitic “s-” is suggested at the first position of the returned list, with the remaining options being also probable. Here is where the existing Romanian BERT model misses the most probable option, due to its non-specific Romanian tokenization. However, the options presented by this model are also very probable, *indicating the fact that our BERT model is still undertrained.*

We verify that our small BERT model can make the same predictions for the predicate even if other words are interposed between it and the subject. We run the sentences “Copiii de la scara noastră vor să se [MASK].” and “Copiii vor să se [MASK].” For these sentences, the five most likely suggestions for the token “[MASK]” are:

Copiii de la scara noastră vor să se vadă .	Copiii vor să se vadă .
Copiii de la scara noastră vor să se întorcă .	Copiii vor să se întorcă .
Copiii de la scara noastră vor să se scrie.	Copiii vor să se ducă .
Copiii de la scara noastră vor să se ducă .	Copiii vor să se știe.
Copiii de la scara noastră vor să se apropie.	Copiii vor să se ia.

We note that the Romanian subjunctive forms “vadă”, “întorcă” and “ducă” appear in both suggestion lists, in the same order, indicating that the prepositional group “de la scara noastră” does not influence the prediction of the predicate.

4 Detecting the Encoding of the Plural Number in Nouns

The experiment aims at determining if there are any dimensions of the resulting encoding of an input token, as produced by a BERT model, that signal the presence of a plural noun. To perform the experiment, we manually selected two sets of plural Romanian nouns with the following properties:

1. They are very frequent in the CoRoLa corpus. Because we work with an intermediary checkpoint of the BERT model, we favour the most frequent nouns, so that the training procedure was given a chance to select them when doing MLM.
2. They are unrelated semantically, and they can only be nouns (no ambiguity with other parts of speech) irrespective of the context. Also, we have no control over the syntactic function of these nouns, so they can appear in any syntactic position (e.g., subjects, objects, or modifiers).

Abiding to these principles, we selected the following nouns in our dataset:

1. neuter, indefinite, nominative-accusative: “animale”, “medicamento”, “principii” and “servicii”.
2. neuter, definite, nominative-accusative: “animalele”, “medicamentele”, “principiile” and “serviciile”.
3. neuter, definite, genitive-dative: “animalelor”, “medicamentelor”, “principiilor” and “serviciilor”.
4. masculine, indefinite, nominative-accusative: “ani”, “autori”, “lei”, “medici”, “oameni”.
5. masculine, definite, nominative-accusative: “anii”, “autorii”, “leii”, “medicii”, “oamenii”.
6. neuter, definite, genitive-dative: “anilor”, “autorilor”, “leilor”, “medicilor”, “oamenilor”.

We are going to randomly select 100 examples for each noun¹⁰, in each group, and compute a Spearman’s rank correlation coefficient between every two dimensions of the encoding vector of the noun¹¹. Each dimension vector has 100 elements (one for each example), and we sort the left dimension vector in ascending order, reordering the elements of the right dimension along with the sorting. For instance, if the left dimension has elements [2, 4, 1, 3] and the second dimension has elements [5, 6, 7, 8] the resulting vectors on which we compute the Spearman coefficient are [1, 2, 3, 4] and [7, 5, 8, 6]. We do the sorting of the first-dimension vector because Spearman correlation detects a synchronous monotonicity of the inputs, and the random sample has no natural order by itself.

The assumption is that if two dimensions of an encoding vector correlate ($R > 0.75$), given the nature of the noun group, correlations along the same dimensions should pinpoint the plural number, if found at the same coordinates for all nouns in one group. The correlation coefficients are plotted in a scatter chart, with dimensions 0–255 (as the hidden size of our BERT model is 256) on the X and Y

¹⁰ From a random sample of 5,000 examples per word form, taken from CoRoLa.

¹¹ We perform this type sampling 50 times for each noun group, to observe clusters of correlations.

axes, with a single figure for all nouns in a group. Figures 2–4 show the scatter plot for noun groups 1–3 above and Figures 5–7 handle noun groups 4–6 above.

Studying Figures 2–4 above, we see clusters of correlations between encoding dimensions 52–60 (X axis) and 86–96 (Y axis) and between dimensions 88–94 and 100–110 that are common to all three figures. There are many more clusters that are common for two out of the three figures, indicating perhaps other commonalities such as syntactic function or some parts of the meaning. There are also clusters of correlations that are visible for just one noun group, as we can see in Figure 4: it seems that the genitive-dative group presents a lot of correlations for dimensions 0–10 with a lot of other dimensions.

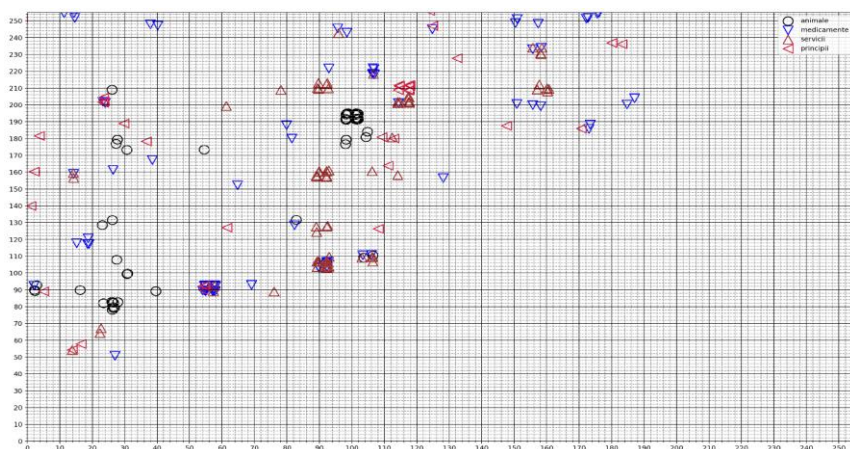


Figure 2. “animale, medicamente, servicii, principii” correlation scatter plot.

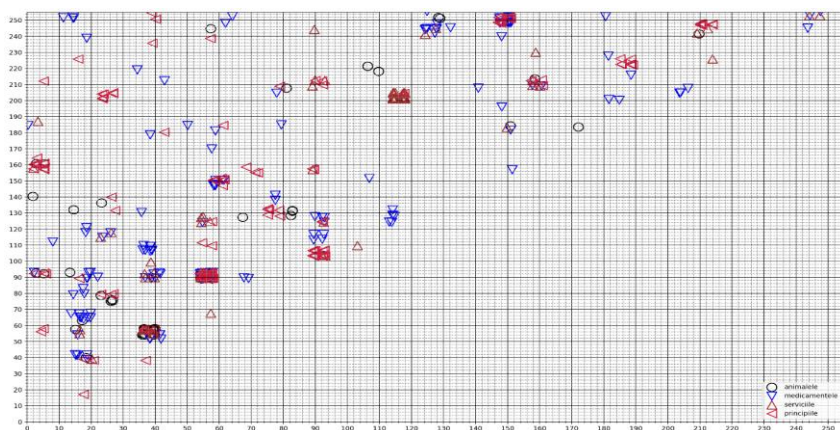


Figure 3. “animalele, medicamentele, serviciile, principiile” correlation scatter plot.

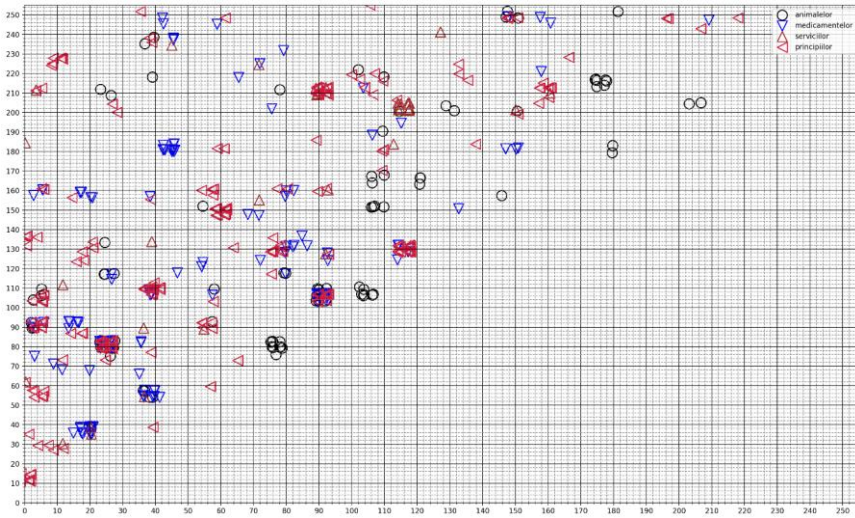


Figure 4. “animalelor, medicamentelor, serviciilor, principiilor” correlation scatter plot.

In Figures 5–7 we observe that the same clusters of encoding dimensions 52–60 (X axis) and 86–96 (Y axis) and dimensions 88–94 and 100–110 are also present in all three figures, making these dimensions good candidates for assuming they encode the plural number in nouns. Looking at Figures 4 and 7 (the genitive noun groups, neuter, and masculine) we clearly see they contain a lot more correlations than the other figures, with many clusters appearing at the same places in the scatter plots. We can hypothesise that the dative-genitive case is more productive from the syntactic and semantic perspectives.

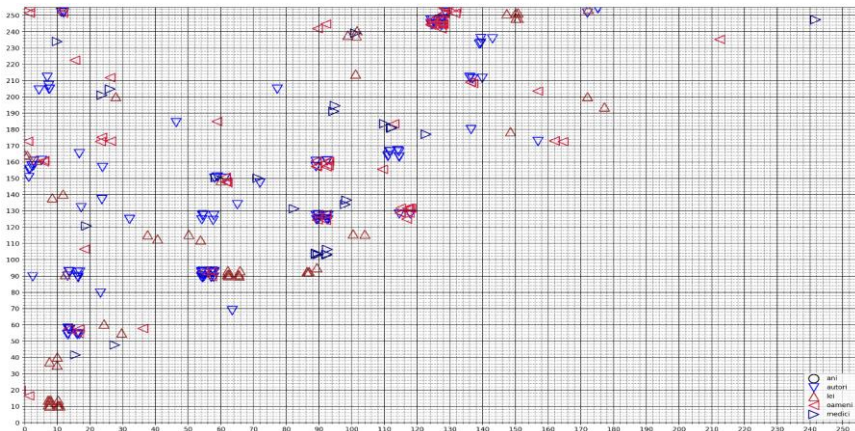


Figure 5. “ani, autori, lei, oameni, medici” correlation scatter plot.

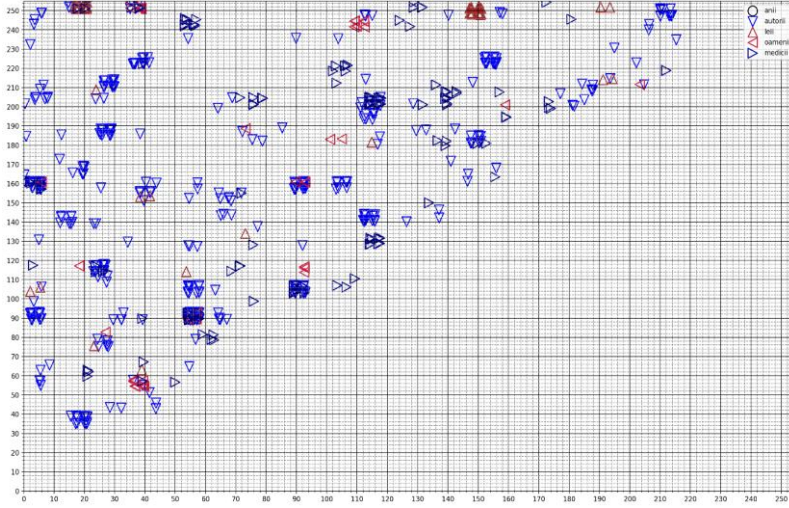


Figure 6. “anii, autorii, leii, oamenii, medicii” correlation scatter plot.

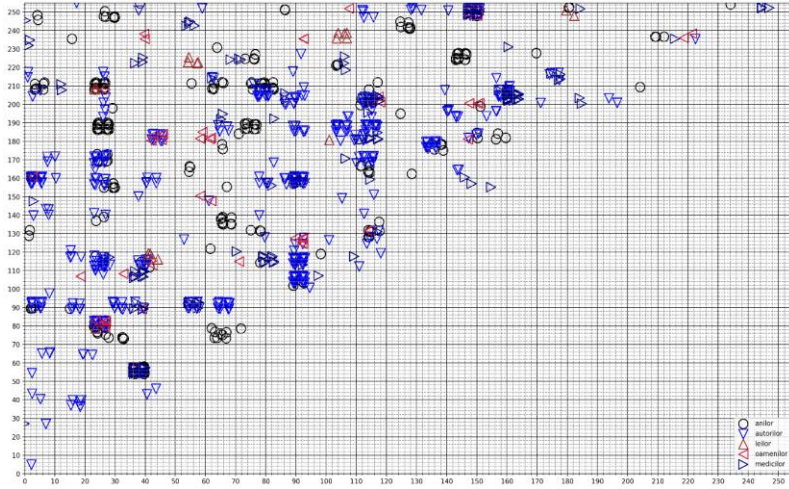


Figure 7. “anilor, autorilor, leilor, oamenilor, medicilor” correlation scatter plot.

5 Conclusions

The present paper introduced a new WordPiece Romanian tokenizer and a BERT model that we train to see how the model learns different linguistic attributes and structures. For a first experiment, we tried to determine which of the dimensions of the encoding representations of input words store the number plural of nouns, and we found that there are common correlation clusters among the six noun

groups that we studied: dimensions 52–60 (X axis) and 86–96 (Y axis) and dimensions 88–94 and 100–110. To further establish that these dimensions indeed encode the plural number, we need to perform a probing experiment (Tenney *et al.*, 2019) in which a classifier is added on top of the word encodings that classifies a token to be “plural”, “maybe plural” (e.g., genitive-dative singular of Romanian feminine nouns) or “not plural”. We can then see which dimensions “fire” when classifying plural nouns and correlate that result with the one we obtained here.

References

- Barbu Mititelu, V., Tufiş, D., Irimia, E., Păiş, V., Ion, R., Diewald, N., Mitrofan, M., and Onofrei, M. 2019. Little Strokes Fell Great Oaks. Creating CoRoLA, the Reference Corpus of Contemporary Romanian. *Revue roumaine de linguistique*, Issue 3. 2019
- Brysbaert, M., Stevens, M., Mandera, P., and Keuleers, E. 2016. How Many Words Do We Know? Practical Estimates of Vocabulary Size Dependent on Word Definition, the Degree of Language Input and the Participant’s Age. *Front. Psychol.* Volume 7 – 2016. <https://doi.org/10.3389/fpsyg.2016.01116>
- Conneau, A., Kruszewski, G., Lample, G., Barrault, L., and Baroni, M. 2018. What You Can Cram Into a Single \$&!#* Vector: Probing Sentence Embeddings for Linguistic Properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126-2136, Melbourne, Australia. Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1198>
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805v2 [cs.CL]*
- Dumitrescu, S. D., Avram, A.-M., and Pyysalo, S. 2020. The Birth of Romanian BERT. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324-4328. Association for Computational Linguistics. Accessed and retrieved online <https://doi.org/10.48550/arXiv.2009.08712>
- Jo, J.-Y. and Myaeng, S.-H. 2020. Roles and Utilization of Attention Heads in Transformer-based Neural Language Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404-3417, July 5-10, 2020. © 2020 Association for Computational Linguistics
- Li, B., Wisniewski, G., and Crabbé, B. 2023. *Assessing the Capacity of Transformer to Abstract Syntactic Representations: A Contrastive Analysis Based on Long-distance Agreement*. Transactions of the Association for Computational Linguistics; 11 18-33. DOI: https://doi.org/10.1162/tacl_a_00531
- OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774v3 [cs.CL]* 27 Mar 2023
- Song, X., Salcianu, A., Song, Y., Dopson, D., and Zhou, D. 2021. Fast WordPiece Tokenization. *arXiv:2012.15524v3 [cs.CL]*, 5 Oct 2021
- Tenney, I., Xia, P., Chen, B., Wang, A., Poliak, A., McCoy, R. T., Kim, N., Van Durme, B., Bowman, S. R., Das, D., and Pavlick, E. 2019. What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations. *arXiv:1905.06316v1 [cs.CL]* 15 May 2019

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2023. Attention is All You Need. *arXiv:1706.03762v7 [cs.CL]*
- Veldhoen, S., Hupkes, D., and Zuidema, W. H. 2016. Diagnostic Classifiers Revealing How Neural Networks Process Hierarchical Structure. In *CoCo@ NIPS*. https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper6.pdf

CHAPTER 5

DISCOURSE AND HUMOUR

ANNOTATING DISCOURSE RELATIONS IN THE ROMANIAN REFERENCE TREEBANK

TUDOR VOICU¹, VERGINICA BARBU MITITELU²

¹*Tudor Vianu National High School of Computer Science, Bucharest, Romania,*
tudor.c.voicu@gmail.com

²*Romanian Academy Research Institute for Artificial Intelligence,*
Bucharest, Romania
vergi@racai.ro

Abstract

We present here the first step taken towards the enrichment of the Romanian Reference Treebank with semantic information. We focus on linking words, which contribute to the cohesion of a text and reveal the logical relations between units of meaning. Starting from a set of Romanian linkers, we annotate their occurrences in the Romanian Reference Treebank using the set of discourse relations defined for the annotation of Penn Discourse Treebank version 3.0. The results of the double annotation are discussed, the types of relations associated with each linker are presented and the way in which this level of annotation is added to the existing resource is explained.

1 Introduction

Natural Language Processing (NLP) has focused on all linguistic levels even for Romanian. It is true, however, that those types of annotation that are independent of or only a little dependent on word meaning have had increasingly better results throughout time (and especially the recent past as an expected outcome of both the hardware and software developments).

In this paper, we focus on a less explored field in Romanian computational linguistics, namely discourse, more specifically on some lexical elements that constitute cohesive devices, i.e., conjunctions.

Cohesion is an important property of a text, which shows that its parts are linked to each other, and that there is a flow of ideas therein. Alongside adverbs, conjunctions¹ are textual connectors, i.e. a means to ensure the formal and semantic connection between discourse segments.

¹ We use the term *conjunction* throughout this paper to refer both to conjunctions and to conjunctive locations.

The level of analysis we adopt here is the sentence. At the syntactic level, conjunctions are coordination or subordination devices, both at the clause (see example (1)) and at the sentence level (see example (2)). In the first example, the conjunction *precum și* “as well as” links two direct objects (*conjunții* and *adverbe*) of the same verb (*vom analiza*), while in the second example the same conjunction is used to link two clauses (*că a participat la eveniment* and *că a făcut o prezentare*) that are subordinated to the same verb (*amintesc*). Our interest lies in the latter use, i.e., in the use of conjunctions as connectors between clauses.

(1) Vom analiza conjuncții, *precum și* adverbe.

“We will analyze conjunctions, as well as adverbs.”

(2) Îmi amintesc că a participat la eveniment, *precum și* că a făcut o prezentare.

“I remember that she participated in the event, as well as that she made a presentation.”

The paper is structured as follows: in Section 2 we present previous work on the annotation of discourse relations in other languages and in Romanian. In Section 3 we introduce the corpus and the inventory of discourse relations we use. Given that the corpus is syntactically annotated, we also explain the principles of annotation that are relevant for the identification and interpretation of conjunctions occurring therein. Section 4 focuses on our work methodology, while the very preliminary results are presented and discussed in Section 5, before envisaging near future work and concluding the paper (Section 6).

2 Related Work

There are two main lines of research that are relevant for our work: on the one hand, there is the emulation the authors of the PDTB principles of annotation have created with respect to developing corpora annotated according to the same principles and using the same inventory of relations. On the other hand, we look into other endeavours of annotating Romanian texts with discourse relations.

2.1 Annotation of Discourse Relations

The creation of the PDTB corpus annotated with discourse relations (Prasad *et al.*, 2019) encouraged other teams to create corpora annotated with the same inventory of relations for other languages. However, such initiatives are not numerous. One such endeavour is presented by Zeyrek *et al.* (2020), who created a parallel corpus made up of TED-talks annotated for 6 languages (English, Polish, German, Russian, European Portuguese, and Turkish) following the PDTB guidelines. A corpus for Lithuanian was also annotated with the PDTB relations (Oleškevičienė

et al., 2023). Feltracco *et al.* (2017) report on the creation of a corpus of Italian annotated only with the contrast relations defined for PDTB.

2.2 Annotation of Discourse Relations in Romanian Texts

One of the preoccupations with the annotation at the discourse level in Romanian corpora had as a result the Quo Vadis Corpus (Cristea *et al.*, 2015). It is a belletrist text annotated with a focus on semantic relations (Cristea *et al.*, 2015). The corpus is a result of a mix of automatic and manual annotation, with over 45,000 annotations belonging to four binary types: *anaphoric*, *affectional*, *kinship*, and *social*. The automatic annotation involves tokens, parts of speech, and lemma layers, while the entity and relation layers (Gifu and Vasilache, 2014) were added manually by trained annotators. The annotated corpus was not made publicly available.

Ștefănescu *et al.* (2020) present the results of annotating two textual connectors, *de altfel* and *de altminteri* in a set of 150 sentences extracted from CoRoLa (Tufiș *et al.*, 2019) with the inventory relations proposed by Crible and Degand (2019). The annotation was made manually by three linguists. It was corpus work, with a purely pragmatic interest, and no annotated data was made available.

Postolea (2018) reports on the linguistic results of the manual annotation of adversative conjunctions (*\dar, iar, însă, ci* “but”) in a set of 200 EuroParl parallel statements, originally in Romanian and translated into English, using PDTB 2.0 Annotation Manual (Prasad *et al.*, 2007). The annotated data is not made available. The work reported here is the closest to our endeavour, due to the inventory of textual relations used, which is an older version of the ones we used.

3 Prerequisites

Before delving into our work, we offer information meant for understanding our approach. We briefly present the corpus used in the annotation process, its structure, the annotation framework and its format (Section 3.1) and we get into more detail concerning the annotation of conjunctions in this corpus (Section 3.2). On the other hand, we overview the inventory of relations used in the annotation (Section 3.3).

3.1 Romanian Reference Treebank

The corpus we are working with for annotating discourse relations expressed by conjunctions is the Romanian Reference Treebank (henceforth, RRT) (Barbu Mititelu, 2018). It is a collection of 9,523 sentences from several genres: legal,

news, fiction, medical, non-fiction, academic, FrameNet translations and Wikipedia, with an unbalanced distribution. The corpus is morpho-syntactically annotated using dependency relations and following the principles of Universal Dependencies² (henceforth, UD) (de Marneffe *et al.*, 2021). It is released biannually, within UD. The corpus is available in CoNLL-U format³, a tabular format with ten columns, storing various types of information: word ID, form, lemma, universal part-of-speech tag (UPOS), language-specific part-of-speech (XPOS), a list of morphological features (FEATS), syntactic head of the current word, the dependency relation the current word establishes with its syntactic head, enhanced dependency graph and any other annotation.

UD principles of syntactic annotation give primacy to content words, which means that dependency relations hold directly between content words, not mediated by function words, which are (always) annotated as leaves in the trees. Content words are nouns, adjectives, verbs (but not auxiliaries) and adverbs, while the rest are function words.

3.2 Conjunctions in Universal Dependencies

Conjunctions are function words, i.e., words that may or may not have a meaning of their own but help clarify the syntactic and/or semantic relation between other words or phrases. Such words are analysed as dependents in UD, they are not heads, irrespective of their status: coordinating or subordinating conjunctions. An example of the way in which coordinating conjunctions are annotated is offered in Figure 1, which contains the representation of the dependency relations in the sentence in example (3): we notice the conjunction *și* “and” attached as a dependent by means of the relation *cc* to the head of the second conjunct, namely the verb *spun* “say”.

(3) Dați-mi voie și vă spun, cuvânt cu cuvânt.

“Allow me and I will tell you, word by word.”

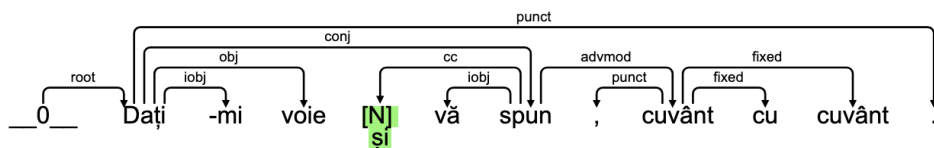


Figure 1. Representation of dependency relations:
function words are always dependents: the case of coordinating conjunctions.

² <https://universaldependencies.org/>

³ <https://universaldependencies.org/format.html>

Subordinating conjunctions are also dependents, as can be seen in Figure 2, which offers a graphical representation of the dependency relations in the sentence in example (4), where the conjunction *deoarece* “because” is a dependent of the head of the subordinate clause it introduces, namely the verb *lovit* “hit”. The syntactic relation specific to subordinating conjunctions is *mark*.

- (4) Unul din ecrane s-a întrerupt deoarece el a lovit un avocat.
 “One of the screens turned off because he hit a lawyer.”

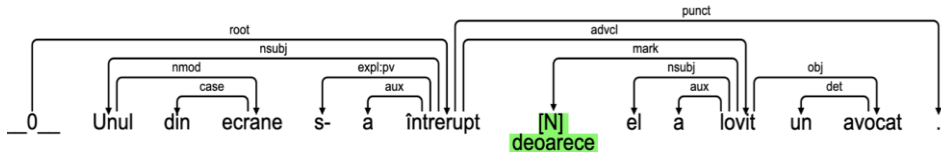


Figure 2. Representation of dependency relations:
 function words are always dependents: the case of subordinating conjunctions.

Conjunctive locutions are treated as multiword tokens, i.e., they are segmented into components and only the first component establishes the relation *mark* with the verbal head, while the other component(s) is/are dependent(s) of the first one and are linked by means of the *fixed* relation: see Figure 3 which offers a graphical representation of the dependency relations in the sentence in example (5): *în*, the first element of the locution *în timp ce* “while”, establishes the *mark* relation with the verb *zbătea*, while *timp* and *ce*, each establishes the *fixed* relation with *în*.

- (5) Echipajul său de opt oameni a pierit de asemenea în timp ce se zbătea să salveze echipajul Santampa.
 “His crew of eight people also died while striving to save the Santampa crew.”

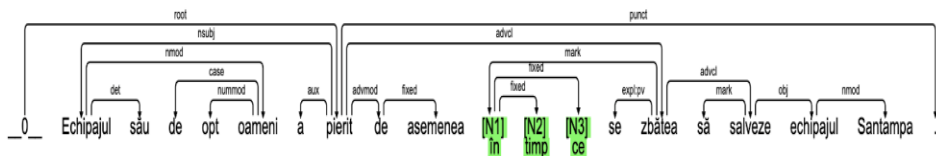


Figure 3. Representation of dependency relations:
 annotation of conjunctive locutions.

Once aware of the way in which conjunctions are annotated within RRT, it is easy to automatically find them in the treebank and extract them. However, there are cases when a conjunctive locution can not be so easily retrieved from the treebank: in example (6) below, there is the conjunctive locution *fără să*

“without”, but its second component (*să*) is also a marker of the conjunctive mood of the verb. Thus, in our analysis of this conjunction, we chose to annotate *să* as a marker of the verb, instead of identifying the conjunctive locution: Figure 4 shows this way of analysis, which is totally agnostic to the locutional status of *fără să* in Romanian traditional grammar. With this new annotation we present here, it becomes possible to record this information, as we show in Section 5.

- (6) El continuase să meargă, fără să se oprească.
 “He had continued walking, without stopping.”

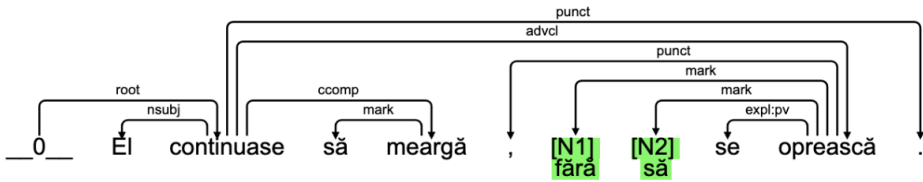


Figure 4. Representation of dependency relations:
 annotation of a conjunctive locution ending in the subjunctive marker *să*.

3.3 PDTB 3.0 Inventory of Discourse Relations

The Penn Discourse Treebank⁴ (PDTB) (Prasad *et al.*, 2019) is a corpus annotated with discourse relations. The texts belong to the journalistic genre: they are articles from *The Wall Street Journal*, thus reflecting a specialised language in over 1 million words, containing 40,600 tokens of annotated relations, making it the largest such corpus available today. The annotation manual used (Webber *et al.*, 2019) was also made publicly available, and therefore investigating discourse relations in the Romanian language is made possible due to its existence.

The inventory of discourse relations coined in the PDTB 3.0 manual are presented in Table 2 there are 4 major relations (Temporal, Contingency, Comparison, and Expansion) subsuming a different number of other relation (see the relations of Level-2 in Table 2), some of which subsume other relations (see those under Level-3). As one can notice, some of these sub-relations and sub-sub-relations are coined in order to account for implicit beliefs (epistemic knowledge) and speech acts associated with the arguments of the respective relations.

Each relation is established between two arguments. An argument is usually a clause, in syntactic terms, i.e., a syntactic unit organised around a verb. For instance, in example (3) there are two arguments: *Dați-mi voie* and *vă spun, cuvânt cu cuvânt*. They are linked by means of the conjunction *și*.

⁴ <https://catalog.ldc.upenn.edu/LDC2019T05>

4 Work Method

4.1 Selection of Conjunctions for Annotation

In the initial phase, we identified the conjunctions within the Romanian language that express relevant discourse relations. Having an extensive set of 122 conjunctions, 51 simple conjunctions, and 71 multiword conjunctions, we analysed their frequency in RRT. We started the annotation with those with the highest numbers of occurrences (see the last column of Table 1). For this version of the paper, we report the results of the annotation of four conjunctions, *în timp ce* “while”, *pentru că* “because”, *după ce* “after” and *înainte de* “before”, but, given that the annotation is in progress, we will be able to offer results for more conjunctions in the very near future.

4.2 Annotation Method

For each occurrence of a conjunction, the two arguments that it links are identified and then a relation from the PDTB hierarchy (Table 2) is assigned to it: only relations of the third level are assigned; when there is no relation on this level, a second level one is chosen. In case the context is insufficient, the annotation is skipped. This is possible because RRT does not necessarily contain contiguous sentences which would offer a larger linguistic context for interpretation.

4.3 Inter-Annotator Agreement

In order to maintain a high degree of accuracy and consistency, each occurrence was annotated by two annotators. This approach is meant to ensure the quality of the results, and facilitates the identification of any discrepancies between annotations, of the difficulties faced in this process.

Table 1 shows, in the second column, the Kappa score calculated for the parallel annotations of the conjunctions. Its values for different conjunctions show that they raise different challenges for the annotators. We noticed that differences mostly occur with relations involving beliefs and speech acts, which are more difficult to recognize (see Section 5).

Table 1. Kappa score for and frequency of occurrence of the annotated conjunctions.

Conjunction	Kappa score	Frequency
<i>în timp ce</i> “while”	0.743	51
<i>pentru că</i> “because”	0.224	54
<i>după ce</i> “after”	0.791	87
<i>înainte de</i> “before”	1.000	35

Table 2. PDTB-3 Sense Hierarchy (Webber *et al.*, 2019).

Level-1	Level-2	Level-3
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE
		SUCCESSION
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND
		ARG2-AS-COND
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND
		ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL
		ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER
		ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST
		ARG2-AS-SUBST

5 Preliminary Results and Discussion

We show the relations expressed by the annotated conjunctions in Table 3. The values presented here are those established by the two annotators after discussing together the cases when they annotated differently. We can see that only one of the annotated conjunctions expresses only one relation, namely *înainte de*, which expresses PRECEDENCE, while the others are used to express more relations. However, there is one value that prevails for each of them: TEMPORAL:SYNCHRONOUS for *în timp ce*, CONTINGENCY:CAUSE:REASON for *pentru că* and TEMPORAL:ASYNCHRONOUS:SUCCESSION for *după ce*. There is also another relation that stands out among the others for each conjunction: COMPARISON:CONTRAST for *în timp ce* and CONTINGENCY:CAUSE+BELIEF:REASON+BELIEF for *pentru că*; a peculiarity of *după ce* is the fact that it cumulates two meanings: there is an important number of contexts in which it has two values: TEMPORAL:ASYNCHRONOUS:SUCCESSION and CONTINGENCY:CAUSE:REASON. The other relations annotated only occur in a few sentences. This polysemous nature of the conjunctions is one of the sources of disagreements between annotators.

Although the discourse relations expressed by these two conjunctions are totally different, we, of course, expect cases when different conjunctions will express the same relation, thus showing synonymy.

As explained above, the RRT is available in CoNLL-U format, a tabular format with ten columns, including the word ID, the syntactic head of the current word, the dependency relation the current word establishes with its syntactic head, and any other annotation needed. In the case of a simple conjunction, the dependency relation will be annotated as *mark*. A conjunctive locution is treated as a multitoken word, as explained above (Section 3.2).

Table 3. The discourse relations expressed by the annotated conjunctions.

Conjunction	Frequency	Annotation
<i>în timp ce</i> “while”	29	TEMPORAL:SYNCHRONOUS
	18	COMPARISON:CONTRAST
	2	COMPARISON:CONCESSION:ARG2-AS-DENIER
	1	COMPARISON:SIMILARITY
	1	insufficient context
<i>pentru că</i> “because”	35	CONTINGENCY:CAUSE:REASON
	10	CONTINGENCY:CAUSE+BELIEF:REASON+BELIEF
	4	CONTINGENCY:CAUSE+SPEECHACT:REASON+SPEECHACT
	3	CONTINGENCY:CAUSE+SPEECHACT:RESULT+SPEECHACT
	2	CONTINGENCY:CAUSE:RESULT

Conjunction	Frequency	Annotation
după ce “after”	48	TEMPORAL:ASYNCHRONOUS:SUCCESSION
	38	TEMPORAL:ASYNCHRONOUS:SUCCESSION and CONTINGENCY:CAUSE:REASON
	1	TEMPORAL:ASYNCHRONOUS:PRECEDENCE and TEMPORAL:ASYNCHRONOUS:SUCCESSION
înainte de “before”	35	TEMPORAL:ASYNCHRONOUS:PRECEDENCE

The annotated discourse relations are added on the tenth column of the CoNLL-U file: the first element of the location gets a numerical identifier followed by a colon and the name of the relation; the other elements only get the same number as the first element and the name of the relation remains unspecified. Additionally, when more conjunctive locutions occur within a sentence, each of them is assigned another number. Figure 5 shows the CoNLL-U format of the sentence in example (7). The annotated locutions are highlighted in green. In their last column, we notice the number of the locution and, on its first component, the name of the relation is also added. The same number assigned to different tokens means that they are components of the same locution.

(7) Articulațiile între șanțurile de plastic se pot scurge, de asemenea - de obicei pentru că gunoiul sau pietrișul s-au adunat între șanț și sigiliu sau pentru că însuși sigiliul s-a STRICAT.

“Joints between the plastic trenches can also leak - usually because debris or gravel has collected between the trench and the seal, or because the seal itself has BROKEN.”

1	Articulație	articulație	NOUN	Ncpry	Case=Acc,Nom Definite=Def Gender=Fem Number=Plur	7	nsobj	-	-
2	între	între	ADP	Spua	AdpType=Prep Case=Acc	3	case	-	-
3	șanțurile	șanț	NOUN	Ncpry	Case=Acc,Nom Definite=Def Gender=Fem Number=Plur	1	mmod	-	-
4	de	de	ADP	Spua	AdpType=Prep Case=Acc	5	case	-	-
5	plastic	plastic	NOUN	Ncms-n	Definite=Ind Gender=Masc Number=Sing	3	mmod	-	-
6	se	sine	PRON	Pc3-a-----w	Case=Acc Person=3 PronType=Pcs Reflex=Yes Strength=Weak	8	expl pv	-	-
7	pot	putea	VERB	Vmip p	Mood=Ind Number=Plur Person=3 Tense=Pres VerbForm=Fin	0	root	-	-
8	scurge	scurge	VERB	Vmnp	Tense=Pres VerbForm=Inf	7	ccomp	-	SpaceAfter=No
9	,	,	PUNCT	COMMA	-	10	punct	-	-
10	de	de	ADP	Spua	AdpType=Prep Case=Acc	7	-	-	-
11	asemenea	asemenea	ADP	Alp	Degree=Pos	10	fixed	-	-
12	,	,	PUNCT	DASH	-	22	punct	-	-
13	de	de	ADP	Spua	AdpType=Prep Case=Acc	22	advmod	-	-
14	obicei	obicei	NOUN	Ncms-n	Definite=Ind Gender=Masc Number=Sing	13	fixed	-	-
15	pentru	pentru	ADP	Spua	AdpType=Prep Case=Acc	22	mark	-	-
16	că	că	SCONJ	Cconj	Polarity=Pos	11	fixed	-	-
17	gunoiul	gunoa	NOUN	Ncmry	Case=Acc,Nom Definite=Def Gender=Masc Number=Sing	22	nsobj	-	-
18	sau	sau	CCONJ	Ccop	Polarity=Pos	19	cc	-	-
19	pietrișul	pietriș	NOUN	Ncmry	Case=Acc,Nom Definite=Def Gender=Masc Number=Sing	22	nsobj	-	-
20	sine	sine	PRON	Pc3-a-----w	Case=Acc Person=3 PronType=Pcs Reflex=Yes Strength=Weak Variant=Short	22	conj	-	SpaceAfter=No
21	au	aux	AUX	Va-3p	Number=Plur Person=3	22	aux	-	-
22	adunat	aduna	VERB	Vmnp-am	Gender=Masc Number=Sing VerbForm=Part	7	advcl	-	-
23	între	între	ADP	Spua	AdpType=Prep Case=Acc	24	case	-	-
24	șanț	șanț	NOUN	Ncms-n	Definite=Ind Gender=Masc Number=Sing	22	obl	-	-
25	și	și	CCONJ	Ccop	Polarity=Pos	26	cc	-	-
26	sigiliu	sigiliu	NOUN	Ncms-n	Definite=Ind Gender=Masc Number=Sing	24	conj	-	-
27	sau	sau	CCONJ	Ccop	Polarity=Pos	34	cc	-	-
28	pentru	pentru	ADP	Spua	AdpType=Prep Case=Acc	24	mark	-	-
29	că	că	SCONJ	Cconj	Polarity=Pos	26	fixed	-	-
30	însuși	însuși	DET	Dclms	Gender=Masc Number=Sing Person=3 PronType=Emp	31	det	-	-
31	sigiliul	sigiliu	NOUN	Ncmry	Case=Acc,Nom Definite=Def Gender=Masc Number=Sing	34	nsobj	-	-
32	sine	sine	PRON	Pc3-a-----w	Case=Acc Person=3 PronType=Pcs Reflex=Yes Strength=Weak Variant=Short	34	expl pv	-	SpaceAfter=No
33	au	aux	AUX	Va-3s	Number=Sing Person=3	34	aux	-	-
34	STRICAT	strica	VERB	Vmnp-am	Gender=Masc Number=Sing VerbForm=Part	22	conj	-	SpaceAfter=No
35	,	,	PUNCT	PERIOD	-	7	punct	-	-

Figure 5. Representation of discourse relations in the RRT CoNLL-U file.

6 Conclusions and Future Work

We have presented here the framework and the first step taken towards the enrichment of RRT with discourse relations from the PDTB 3.0 hierarchy. The work is only at the beginning; thus, the reported results are scarce. The data can give valuable insights into the behaviour of Romanian conjunctions with respect to the discourse relations they express, to their default readings, and also to possible misinterpretations (in cases revealed by the double annotation we are carrying out). Comparisons between conjunctives considered equivalent in different languages (i.e., those for which such annotations do exist) will thus be possible and can help the translation process and text interpretation.

References

- Barbu Mititelu, V. 2018. Modern Syntactic Analysis of Romanian. Clasic și modern în cercetarea filologică românească actuală, pages 67-78, Iași, Romania.
- Crible, L. and Degand, L. 2019. Domains and Functions: A Two-Dimensional Account of Discourse Markers. *Discours*, 24:3-35.
- Cristea, D., Gifu, D., Colhon, M., Diac, P., Bibiri, A.-D., Mărănduc, C., and Scutelnicu, L.-A. 2015. Quo Vadis: A Corpus of Entities and Relations, pages 505-543. Springer International Publishing. In N. Gala, R. Rapp and G. B. Enguix (eds.): *Language Production, Cognition, and the Lexicon*, Springer International Publishing Switzerland.
- Cristea, D., Gifu, D., Niculiță, M., Pistol, I., and Sfirnaciuc, D. 2016. A Mixed Approach in Recognising Geographical Entities in Texts. *Linguistic Linked Open Data*, D. Trandabăț and D. Gifu (eds.), Springer-Verlag Berlin Heidelberg, pages 49-64.
- de Marneffe, M. C., Manning, C. D., Nivre, J., and Zeman, D. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255-308, 07.
- Feltracco, A., Magnini, B., and Jezek, E. 2017. Contrast-Ita Bank: A Corpus for Italian Annotated with Discourse Contrast Relations. *Proceedings of the 4th Italian Conference on Computational Linguistics CLiC-it 2017*, Rome, Italy, pages 159-164.
- Gifu, D. and Vasilache, G. 2014. A Language Independent Named Entity Recognition System. *Proceedings of the 10th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Mihaela Colhon, Adrian Iftene, Verginica Barbu Mititelu, Dan Cristea, and Dan Tufiş (eds.), “Alexandru Ioan Cuza” University Publishing House, Iași, pages 181-188.
- Oleškevičienė, G. V., Karaciejūtė, V., and Gulbinskienė, D. 2023. Lithuanian Discourse Markers and Their Relations in a Multilingual Corpus. *Sustainable Multilingualism*, 22(1):258-272.
- Postolea, S. 2018. A Particle Filter Algorithm for Bayesian Word Segmentation Using Annotation to Identify Connective Meanings in a Multilingual Environment. *Romanian and English Contrast Markers in a Parallel Corpus. Cross-Linguistic Discourse Annotation: Applications and Perspectives. TextLink2018*, Toulouse, France, pages 107-113.

- Prasad, R., Miltsakaki, E., Dinesh, N., Lee, A., Joshi, A., Robaldo, L., and Webber, B. 2007. The Penn Discourse Treebank 2.0 Annotation Manual. University of Pennsylvania.
- Prasad, R., Webber, B., Lee, A., and Joshi, A. 2019. Penn Discourse Treebank Version 3.0. Linguistic Data Consortium.
- Tuفی, D., Barbu Mititelu, V., Irímia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., and Onofrei, M. 2019. Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus of Contemporary Romanian. *Revue Roumaine de Linguistique*, 64:227-240.
- Webber, B., Prasad, R., Lee, A., and Joshi, A. 2019. The Penn Discourse Treebank 3.0 Annotation Manual. Philadelphia, University of Pennsylvania, 35:108.
- Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., and Ogrodniczuk, M. 2020. TED Multilingual Discourse Bank (TED-MDB): a Parallel Corpus Annotated in the PDTB Style. *Language Resources Evaluation*, 54:587-613.
- Ştefănescu, A., Postolea, S., and Barbu Mititelu, V. 2020. The Romanian Discourse Markers *de altfel* and *de altminteri*. Patterns of Use and Core Functions. *Revue Roumaine de Linguistique*, 65:307-322.

THIS MAKES ME LAUGH. DOES IT MAKE YOU LAUGH, TOO?

DANIELA GÎFU^{1,2}, DAN S. STOICA³

¹*Faculty of Computer Science, Alexandru Ioan Cuza University,
Gen. Berthelot St. 16, 700483, Iași, Romania*

²*Institute of Computer Science, Romanian Academy, Codrescu St. 2, 700481,
Iași, Romania*

`daniela.gifu@iit.academiaromana-is.ro`

³*Faculty of Letters, Alexandru Ioan Cuza University, Carol I Blvd. 11, 700506,
Iași, Romania*

`dstoicaro@gmail.com`

Abstract

As a general rule, humor is supposed to facilitate social interaction, becoming a way to solve a communication crisis. One could also notice that humor is always at the expense of someone, of some creature, or even of something, which could influence the course of relationships in society. People react differently to humor and there are different causes to it. In fact, the explanation for such different reactions could be individuals' natural constitution (not all brains are functioning in the same way), the acquired knowledge of each person, their social abilities, but it can be the effect of the cultural pressure too. In the end, we are what our respective cultures make us be. Up to a point! Therefore, humor provides insight into how humans process the language, which is a creative, alive, complex endeavour, not just applying an adaptable subset of standard sentences, as well as how they cope with different contexts. Considering linguistic humor theories, we propose a methodology of humor comprehension based on semantic interpretation, within different cultural contexts. Since the early 1990s, computer scientists and linguists have been raising the issue of recognizing, even of creating jokes using machine learning or/and deep learning algorithms. This paper aims to suggest such a perspective over the topic.

Keywords: humor reception, cultural patterns, stereotypes, semantics

1 Introduction

In our life, humor creates an essential environment for communication. In fact, it is an intelligent tool in human interaction (Chen & Soo, 2018). Being closely

related to figurative language, recognizing and understanding humor is one of the most challenging tasks in natural language processing (NLP) (Samson & Gifu, 2021). What will interest us most is the fact that, even though humor is present in all human cultures, it seems to have different manifestations in each of them (Jiang *et al.*, 2019).

The legitimate question of this survey is: *How does culture influence the perception of humor?*

The user-generated-content demands a proper way to recognize specific information that is relevant to detect humorous tendencies (Mao & Liu, 2019). Our approach aims to highlight the connection between culture and humor. In fact, this survey is part of a research on a broad range of topics in humor studies (Bernab  , 2023; Dore, 2022; Gironzetti, 2022).

The paper is structured as follow: Section 2 presents a short background of a number of humor studies in order to have a realistic starting point about how culture influences the construction of humor, while Section 3 describes a narrative humor recognition method focused on examples where humor is related with stereotypes and prejudices from different cultures, significant for future research based on the selection and training of the machine learning models, along with the evaluation metrics used to assess its performance, before drawing some conclusions in the last section.

2 Humor and Culture

Like most linguistic manifestations, humor is produced by people according to the culture they belong to, across all societies (Bryant & Bainbridge, 2022). But despite this fact, there is not enough study conducted on the examination of structural elements of humor-related to cultures (Lau *et al.*, 2023).

Here, we present existing research examining the perception of humor across cultures. And yet, one should not oversee the fact that communication, in general, is always about a cultural model. We couldn't really capture what happens in a communication instance while ignoring that communication is culture sensitive.

2.1 Defining Culture

Tricky! Let's try a definition which could also help in distinguishing between many kinds of cultures.

Culture could be a particular way of using signs and symbols to signify things to one another, but also all the accepted (and thus shared) customs and traditions, values and taboos, *doxa* and *topoi* as well as a specific (almost unlimited) list of situations of communication (where a situation of communication should be taken as the occurrence of a social practice). If we can accept the existence of common

characteristics for a given population, we should understand that cultures are imprinted in each and every one of us humans, and this makes us resemble the people around us as well as being different from others.

Every society is defined, among other things, by its mind formatting institutions, those which take care of social reproduction (on this topic, see also: Stoica 2015). This is how mentality is built as the most time-resisting shared knowledge of the world and of how life is. Cultures are broken down into subcultures which contain sub-subcultures and so on. Or this makes us resemble other members of our culture to different degrees and from different points of view.

Let's take us, the Romanians: we use the same signs to convey approximately the same meanings (with differences relying on different levels of language knowledge or the knowledge of other codes), we share (almost!) the same values and the same traditions. Almost! A value like *family*, for example, has been the subject of hot debates and it still keeps us divided.

The problem with culture and mentality is that they imply mind shortcuts, like prejudices and stereotypes, and this gets us back to humor.

2.2 Culture Difference in Humor Reception

Cultures and subcultures will prove to be different when it comes to creating or enjoying humor. But it is certain that humor is a form of communicative behaviour consisting in that the transmitter expects an immediate particular reaction (a certain type of response) in the receiver. Humor becomes ironic (Reboul, 1980) when the playful is substituted by offensive intentions and the ironic manifestation attracts some limitations (Reyes *et al.*, 2012). Pushing things over a limit, irony becomes sarcasm, which is also a form of humor but from the social point of view it is rougher.

A large part of popular jokes is built starting from prejudices and stereotypes (concerning nations, regions, groups, individuals): we all know that Germans are strict, French are flamboyant, British are formal, but also Romanians from the South seem to know that those from Moldavia are lazy and those from Transylvania are slow, peoples from Northern Europe think that those from the South are lazy and unreliable, while those from the South think that Northerners are frozen. A joke about a stingy character is always about Scots, not only in England (even if the reality proves that Scots are much more generous than people in England).

Basically, humorous statements can produce unexpected, sometimes contrary, effects in the hearer – cohesion (Wennerstrom, 2011) and exclusion (Mathias, 2022) – and, as such, an objective evaluation becomes essential.

Interesting enough, jokes can also come from attacking prejudices and stereotypes. As prejudices and stereotypes do not exist as such but as sources of people's behaviour, one could laugh at the expense of someone else who proves to be the slave of some prejudice or stereotype. And there is more: jokes could come from political correctness fighting prejudices and stereotypes, and also from mocking the political correctness which fights prejudices and stereotypes. We are all different from one another and this makes us have different relations with humor.

There is more. The emotions experienced through humor rely on and enhance the complicity between those who take part in the communicative interaction, the foundation of which lies in the shared mentality as well as in the shared way in life. Would someone be a member of a society where they value violent shows, where violent death of different creatures occurs? A joke about the last spasms of a dying creature would bring a smile on their face and a pleasant emotion in their heart. The same kind of shows could be horrific for someone from another culture, where the distance in time from those epochs which knew public executions of humans is valued and hailed. This discussion on the deep explanation of good mood and laughter will be addressed in a further study.

3 Narrative Humor Recognition Method

A few years ago, we explored the applicability of computational methods to the recognition of Adjectival Noun Phrases (ANPs) expressing humor in religious speech (Gifu *et al.*, 2014). This topic has attracted significant attention in recent years, something that could be noticed due to the increasing number of workshops (e.g., SemEval-2017 Task 6: HashtagWars: Learning a Sense of Humor; SemEval-2020 Task 7: Assessing Humor in Edited News Headlines or SemEval-2021 Task 7: HaHackathon: Detecting and Rating Humor and Offense). Our research into detecting humor and offence has reached a high score in performance accuracy, then it is labelled as a SOTA (Samson & Gifu, 2021).

3.1 Resources

The study of humor through the use of language should necessarily be approached in an interdisciplinary way connected to computational linguistic methods. To build a specific corpus, we monitored social media in order to stock a collection of humorous writings circulating among Romanians (e.g., Brainly.ro). We picked up from this collection a few humorous expressions that were interpreted qualitatively in order to develop a digital recognition method based on deep learning.

3.2 Semantic Approach to Humor Across Cultural Barriers

This study seeks to analyse semantically ways of creating humor in different cultures. We try to capture the way humor is created semantically with the help of denotation, connotation, collocation, and figurative language. The final target of the study is to analyse the role of culture in discerning humor in social media and how to detect the culture an utterer belongs to starting from their jokes.

Some cultures when it comes down to humor

For example, one could learn from Kate Fox, a well reputed English anthropologist, that “while there may indeed be something distinctive about English humor, the real ‘defining characteristic’ is the value [they] put on humor, the central importance of humor in English culture and social interactions”.

And now we quote: “In other cultures, there is ‘a time and a place’ for humor; it is a special, separate kind of talk. In English conversation, there is always an undercurrent of humor. We can barely manage to say ‘hello’ or comment on the weather without somehow contriving to make a bit of a joke of it, and most English conversation will involve at least some degree of banter, teasing, irony, understatement, humorous self-deprecation, mockery or just silliness. Humor is our ‘default mode’, if you like: we do not have to switch it on deliberately, and we cannot switch it off. For the English, the rules of humor are the cultural equivalent of natural laws – we obey them automatically, rather in the way that we obey the law of gravity”.

The secret seems to be the innate capacity of the English of being “more acutely sensitive than any other nation to the distinction between ‘serious’ and ‘solemn’, between ‘sincerity’ and ‘earnestness’.

Why should we pay attention to these findings of Kate Fox?

For we need to understand how difficult it is to separate humor from other linguistic acts of the English, while they themselves make no distinction. Those who fail to understand this will never really understand English. Your English may be impeccable – says K. Moss – but you will never feel or appear at home in conversation with the English unless you come to grab the sense of the distinctions mentioned above: serious vs. solemn, sincerity vs. earnestness.

So, what are we analysing? Maybe, should we have a look at the humorous reactions to solemn or earnest declarations of others?

It was the case after the bombings of 2005 in London. After the English-speaking world (mainly the Americans) had expressed their support for the inhabitants of the British Capital with phrases like ‘We are all Londoners today’, the people of London replied: ‘If you are all Londoners today, that’s eight quid each for the congestion charge’.

Then, the irony, the understatement, the humorous self-deprecation are the most frequent forms of expression in English conversations, even when serious matters are discussed seriously.

Another British humor theorist, Richard Wiseman – evoked by D. H. Monroe, in his *Argument of Laughter* (Monroe, 1951) –, found that there are four archetypes that keep recurring in British jokes. These would be: (1) someone trying to look clever and falling into ridicule; (2) husbands and wives not loving each other; (3) doctors being insensitive about imminent death of their patients; (4) God making a mistake.

All these themes fall under the incongruity theory, embraced under different forms by Kant, Schopenhauer, Kierkegaard and others. This theory states that incongruity is some sort of unusual juxtaposition of events or ideas. In this case, humor would come from some incongruity between expectations and reality. The incongruity theory of humor contains no evidence of human inferiority.

But the same author – D. H. Monroe – reaches the conclusion that there are two more Themes, resulting in two theories: the relief theory (which seems to be more of a psycho-physiological order) and the superiority theory (which interests our present approach).

Present since Plato and Aristotle, the superiority theory could be at work in comedies where the laughter comes from the feeling of being superior to the unfortunate characters who suffer the misfortune of being deluded about their own wisdom. Plato goes further in his *Republic* (Plato, 2018) saying that any story presenting persons or gods as “overcome by laughter” should be banned. As for Aristotle, he finds that laughing at others' ugliness indicates a baseness of the soul.

The same position is expressed by Thomas Hobbes in the 17th century and later by Charles Baudelaire in the 19th century. This latter considers laughter at the misfortune of the others as diabolical and mean.

In comedies, as in all the plays, the characters stand for paradigms, categories of human beings. All these observations tend to promote the idea that there are categories of people who feel superior to other categories of people. They have some sudden conception of eminence in themselves, as Hobbes put it.

Over the Channel, the situation is quite different. The sparkling French spirit springs humor all the time, but not permanently. The French have the sense of the solemn and of the earnestness. For them, there is a time and there is a place for humor, thus they know how to be socially adequate. This makes it easier to capture French humor within the infinite linguistic interactions taking place all the time. A recurrent theme in French humor is of a not politically correct nature. They always had ‘traditional enemies’: first, the English (for centuries), then the Germans (in the 19th and the 20th centuries), and they have consistent series of jokes at the expense of both those nations, some of which entered the folklore

(like the phrase accompanying almost any glass of wine, mainly in the countryside: ‘encore un que les Boches n’auront pas’ (another one the Krauts won’t get)).

For Americans, the French aren’t paying attention to personal hygiene (“What are you, French?”) would be the reproach addressed to someone who didn’t shower from some time), while the English have twisted minds. For the French, the Americans are too young a nation to have their own culinary art.

Stereotypes

Humorous writings rely on these mind shortcuts. The series *The Xenophobe Guide*¹ (Moroiu, 2009), or *The Notebooks of Major Thompson* by the French author Robert Daninos (Daninos, 1986) are just examples.

Going backwards, one could find the prejudices and the stereotypes on which jokes were built. The well-known patterns of jokes circulating among Romanians which start with “Once there was a French guy, a Russian and a Romanian” (where the respective nationalities are variables) makes us think that we are in the presence of occurrence of some stereotypes describing the respective nations.

Living in a weak state, with weak institutions, us, Romanians, value the power of being well connected. Nothing is impossible if you know a guy who knows a guy who could get the job done. Impregnated as we are with this kind of mentality, we simply laugh at the German, or the Swiss way in life: being serious, strict, and obeying to the rules is ridiculously funny and jokes come easily from there. On the other hand, jokes about the loose approach of the facts of life in the Balkans are well appreciated in Western Europe. One could notice the same intercultural perception over the subject “work”: Romanians would laugh at the “workaholic” people from, say, the Netherlands, while the Dutch make jokes about the irresponsible naivety of those who believe that things can be done without effort.

As for the subcultures, there are jokes circulating in Switzerland about the inhabitants of one of the cantons who seem to take showers rarely (they say that birds flying over that canton use just one of their wings, as they need the other one to cover their noses). Similar jokes are appreciated in certain regions of Romania, to laugh at the expense of the inhabitants of other regions susceptible of bathing rarely.

Then, there are professional subcultures. People make jokes at the expense of engineers (prejudicial assumption: they can’t talk), while the engineers make jokes at the expense of people outside their profession who lack technical abilities).

¹ Original English title: The xenophobe's guide to the French

Interestingly enough, the paradoxical perception Romanians have of themselves, globally, contains characteristics like heroic behaviour, love for their country, honesty, but also lack of seriousness, tendency to deceive the others, even tendency to steal. There are no jokes on the first category, but there are numerous jokes on the second category (example: the test three people were submitted to: to accept to be locked up in an empty room, having nothing else than three small metallic balls; after a while, the jury could assess the results of the test: the German guy had made an installation for cleaning the air, the French guy had made an artistic structure, but the Romanian had nothing to show and the three balls had disappeared).

From this point of view, it seems that one of the most adequate readings on the Romanians would be *The adventures of Păcală* (Dulfu, 2000), which provide good reasons to laugh, in an auto ironic view of ourselves.

4 Conclusions

Given the purpose of this pilot project, we have focused mostly on understanding the link between culture and humor. As stated before, humor is always at the expense of someone, of some creature, or even of something. Stressing a feature, a character trait, a specific way of being seems to be the technique, with this particular characteristic that it has to be something out of the ordinary and necessarily not that good (with respect to the standards of the culture of the other). A funny way of walking, of talking or of wearing one's glasses, something strange in one's proportions, a funny way a dog has in saluting his owner, a nice but funny way a cat has in taking part in humans' daily routine, some disproportion one cannot not notice in the furniture disposition in someone's apartment, all these are but a few examples of possible subjects for jokes.

Do we all notice the above (possibly) funny elements around us? Do we all find they are funny, worth joking about?

No and no! Some of us may notice strangeness in people, creatures, or things, but they may not see an opportunity to make a joke or even accept that these kinds of situations could be subjects to joke about. Others may not notice anything strange – or laughable for that matter – in the world around. The explanation for such different reactions could be the natural differences between individuals, but it can be the effect of the fact that each and every one of us is part of their respective culture, and all a culture consists of is unavoidable for the individual.

So, yes, we are the product of our respective cultures. Not entirely, of course, but largely and involuntarily.

References

- Bernabéu, E. L. 2023. *The Pragmatics of Humour in Interactive Contexts*. [Pragmatics & Beyond New Series, 335] 2023.vi. © John Benjamins. page 239
- Bryant, G.A. and Bainbridge, C.M. 2022. Laughter and Culture. *Phil. Trans. R. Soc. B*, DOI: 10.1098/rstb.2021.0179.
- Chen, P.-Y. and Soo, V.-W. 2018. Humor Recognition Using Deep Learning. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2 (Short Papers), New Orleans, Louisiana. ACL, pages 113-117.
- Daninos, P. 1986. *Les carnets du major W. Marmaduke Thompson: découverte de la France et des Français*, Paris: Hachette.
- Dore, M. 2022. *Humour in Self-Translation*. [Topics in Humor Research, 11] 2022.xi, © John Benjamins, 278 pages
- Dulfu, P. 2000. *Isprăvăile lui Păcală*, Cartex. Bucuresti.
- Gironzetti, E. 2022. *The Multimodal Performance of Conversational Humor*. [Figurative Thought and Language, 13] 2022.xix, © John Benjamins, 235 pages
- Gifu, D., Scutelnicu, L.-A., and Cristea D. 2014. Humour and Non-humour in Religious Discourse. *Proceedings of the 2nd Workshop on Language Resources and Evaluation for Religious Texts* co-hosted at LREC'2014. Language Resources and Evaluation Conference, Claire Brierly, Majdi Sawalha, Eric Atwell (eds.), Reykjavik (Iceland), pages 30-35.
- Jiang T., Li, H. and Hou Y. 2019. Cultural Differences in Humor Perception, Usage, and Implications. *Front. Psychol.* 10:123. DOI: 10.3389/fpsyg.2019.00123.
- Lau, C., Swindall, T., Chiesi, F., Quilty, L.C., Chen, H.C., Chan, Y.C., Ruch, W., Proyer, R., Bruno, F., Saklofske, D.H., and Torres-Marín, J., 2023. Cultural Differences in How People Deal with Ridicule and Laughter: Differential Item Functioning between the Taiwanese Chinese and Canadian English Versions of the PhoPhiKat-45. *European Journal of Investigation in Health, Psychology and Education*, 13(2):238-258. DOI: 10.3390/ejihpe13020019
- Mao, J. and Liu, W. 2019. A BERT-based Approach for Automatic Humor Detection and Scoring. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)* co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019), Bilbao, Spain, September 24th, 2019.
- Mathias, M.C. 2022. *Humor and Offense Speech Classification and scoring using Natural Language Processing*, Master Thesis in Data Science, Iscte – Instituto Universitário de Lisboa, Portugal, October 2022, accessed online and retrieved from https://repositorio.iscte-iul.pt/bitstream/10071/26655/1/master_marcelo_custodio_mathias.pdf
- Monro, D. H. 1951. (Sprott, W.J.H. 1952). Argument of Laughter. By D. H. Monro. (Melbourne University Press and Cambridge University Press. Price 21s.), *Philosophy*, 27(103):372-373. DOI: 10.1017/S0031819100034434.
- Moroiu, M. 2009. *Ghidul xenofobului: francezii / Nick Yapp, Michael Syrett*; traducere din limba engleză: Mihai Moroiu. București, Nemira (original English title: The xenophobe's guide to the French).

- Plato. 2018. *The Republic*, edited by G. R. F. Ferrari; translated by Tom Griffith, Cambridge University Press.
- Reboul, O. 1980. *Langage et idéologie*, Paris, Press Universitaires de France.
- Reyes A., Rosso P., and Buscaldi D. 2012. From Humor Recognition to Irony Detection: the Figurative Language of Social Media. *Data & Knowledge Engineering* 74:1-12
- Samson, M. and Gifu, D. 2021. FII FUNNY at SemEval-2021 Task 7: HaHackathon: Detecting and Rating Humor and Offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, ACL, pages 1226-1231.
- Stoica, Dan S. 2015. Mind Formatting Institutions. *TrajEthos*, 4(1), 91-109, 2015.
- Wennerstrom, A. 2011. Rich pitch: The Humorous Effects of Deaccent and L+H* Pitch Accent. *Pragmatics & Cognition*, 19 (2): 310-332. Available online from <http://dx.doi.org/10.1075/pc.19.2.07wen>

DETECTION OF MEANING RELATIONS IN ROMANIAN WORD DEFINITIONS

CĂLINA-GEORGIANA CIOCOIU

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

calinaciocoiu@gmail.com

Abstract

Most of the difficult problems that might appear in natural language processing are connected to semantics; not only because it is hard to train the machine to the perfection of the human brain, but also due to the lack of consistency in the way we define these types of subjects. This study aims to increase the accuracy for identifying semantic relations (synonymy, hypo-hyperonymy, and meronymy): our semasiological approach is using the definition datasets, by exploring The Thesaurus Dictionary of the Romanian Language in electronic format (eDTLR).

Keywords: semasiology, NLP, hyponymy, synonymy, meronymy, annotation, formal expressions.

1 Introduction: Linguistic approach of meaning relations

Dealing with the meaning of words is one of the most common difficulties the machines and their developers face nowadays in almost every NLP (Natural Language Processing) activity. In their own way of processing meanings, humans are not fully aware of how they associate a word with what it expresses. When we face a new, unknown word, we consult lexical resources (dictionaries) that might help us define it; what about the machine? How can we explain to machines what are the semantic relations inside the language based on dictionaries?

Most of the studies from the past years and from this domain focus on a particular aspect that is not always visible from the beginning: meanings in context. This is a subject of case studies (Garcia, 2021), disambiguation tasks (Hauer *et al.*, 2021; Dascalu and Gifu, 2015), and semantics in discourse relations (Reig Alamillo *et al.*, CL 2023; Khan *et al.*, 2022). When we think about WordNet, we already know that it “can be said to be a dictionary based on psycholinguistic principles” (Miller *et al.*, 1993); in other words, it shows words that humans use in different situations. Regarding this, it is important to understand that this lexicon was created using a semantic approach, which involves studying the

meaning of words and their relations inside the language; however, there are studies that analyse the synonyms in a more systematic way by studying definitions (Popović, 2023).

In our study, we will try to define and identify the semantic relations (synonymy, hypo-hyperonymy, and meronymy) using a semasiological approach. Semasiology studies the meaning of signs semi-independent of their usage in language. This part of lexicography uses context as a fact to define the change of meaning on itself (Bahnaru, 2009). For better understanding of the differences between semantics and semasiology, let's say we want to analyse a polysemantic word. If we have a semantic approach, we will identify in which cases we can use each meaning that word has (Bahnaru, 2009). On the other hand, if we decide to conduct a semasiological analysis, we will focus on the meanings by themselves—what are the common points of the meanings, what are the differences between the definitions, what are the meaning particularities for each case?

In order to avoid confusion from now on regarding our approach during the study, instead of naming synonymy, hypo-hyperonymy, meronymy “semantic relations”, we will use the term “meaning relations”; this one is closer to the analysis methods and specific particularities of semasiology.

2 Proposed Method

For extracting as many meaning relations as possible, we should use the linguistic approach to our advantage. Semasiology is often used in defining words and lexicographic material creation. Reversing this process, we try to directly extract the other partner of the relation from definitions.

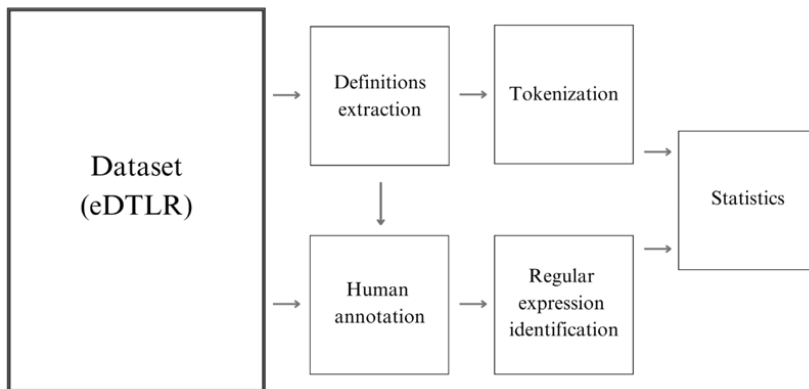


Figure 1. Architecture of the proposed method.

We focus on studying the form of the definitions for a single volume of the dictionary, observing patterns based on human annotation, and generalising certain regular expressions. Consequently, we apply the patterns on the tokenized definitions and extract automatic annotations, based on their type of relation with the defined word.

3 Data Acquisition & Definitions Extraction

For our work of inferring meaning relations from definitions, similar with the datasets selected for similar studies one of the most valuable lexical resources of Romanian language is The Thesaurus Dictionary in electronic form: eDTLR¹. Even if the place of synonyms inside dictionary definitions is not as productive as the appearance of meronyms and hypo-hyperonymy descriptions for their users (Ndlovu, 2009), we still consider this dataset suitable for our study.

For the reason that eDTLR sums up 37 volumes created by different teams in different times (as we discover in the peritexts²), we decided to work with the lexical material of only one: the XI.1 Volume of the Thesaurus Dictionary of Romanian Language (DLR XI.1, 1978). It is one of the best written volumes and it was used as a standard or example.

Moreover, we decided to take into consideration only the nouns, based on the different ways to define different parts of speech in the dictionary. To identify the meaning relations inside the definitions, and to observe how different meaning sets are created, we extracted the words and their definitions from the .xml files.

A necessary way of proceeding here was to not change/clean the data, given the fact that each word, abbreviation, punctuation, and formal aspect represents a part of the definition itself; due this, in the preprocessing part, we just tokenized each extracted item.

4 Data Annotation & Regular Expressions Identification

In order to achieve the best performance and be aware of the importance of each aspect, we were helped by a group of annotators specialised³ in philology.

We have created an Annotation Guide⁴ which contains definitions of the meaning relations, defining their logical type, as follows: synonymy – equality relation, hyperonymy – hierarchical relation, and meronymy – relation of

¹ <https://gitlab.com/solirom-cflr/data/edtlr/-/tree/main> (limited access)

² <https://dlr1.solirom.ro/>.

³ They have at least a BA in Philology, and all of them took linguistics courses.

⁴ The Annotation Guide was written in Romanian, and it is accessible online here.

belonging. The annotators were asked to mark the pair of the given word, found in the definition, and also other seme⁵, using following the markers for:

- "\$" – synonymy;
- "#" – hyperonymy;
- "%0" – meronymy;
- "*" – features;
- "&" – when the annotator is not sure what marker from above to use.

27.	<u>ŞĂBĂCIT</u> = Şăbăceală\$
28.	<u>ŞABANĂ</u> = #Haină# *scumpă*, *aşemănătoare cu un anterior*, pe care *o purtau mai ales vornicii, aprozii şi copiii de casă* *la curţile domneşti*.
29.	<u>ŞABANĂ</u> = #Haină# *lungă* şi *largă*.
30.	<u>ŞABĂR</u> = #Unealtă# folosită *la răzuirea lemnului* sau *a unor piese de metal*.
31.	<u>ŞABĂR</u> = #Dispozitiv# folosit *pentru curăţirea depunerilor de parafină* *de pe pereţii interiori ai ţevilor* *de extracţie* *ale sondelor*.
32.	<u>ŞABĂR</u> = #Dălţiţă# folosită *la lucrări de gravare* *în lemn*, *în metal* etc. sau *la operaţiile de corectare* *a gravurilor, a plăcilor de stereotipie* etc.
33.	<u>ŞABĂS</u> = #Ultimă# #zi# %a săptămînii%, care este *zi de sărbătoare*, *la evrei*. V. \$s a b a t\$.
34.	<u>ŞABĂS</u> = #Reuniune# *intimă*; \$petrecere\$.
35.	<u>ŞABDEAL</u> = #Unealtă# care *se foloseşte la răzuirea doagelor murdare ale butoaielor* (Dobra- Deva).
36.	<u>ŞABLÓN</u> = #Model# în *mărime naturală* *al unei piese*, care *serveşte la ghidarea unei unelte* *pentru prelucrarea* *prin aşchiere* ori *prin deformare*, *pentru limitarea unor porţiuni de suprafaţă pe care se aplică o vopsea* etc. V. \$t i p a r\$.
37.	<u>ŞABLÓN</u> = #Formulă# *stereotipă*, #repetare# *mecanică*, *lipsită de interes şi de spirit inventiv, a unui model*. V. \$c l i ş e u\$.
38.	<u>ŞABLÓN</u> = #Instrument# cu ajutorul căruia *se controlează* sau *se măsoară dimensiunile sau unghiurile unei piese* *în curs de fabricaţie*.
39.	<u>ŞABLÓN</u> = #Instrument# cu care *se reglează aparatele sau mecanismele* din *industria textilă*.

Figure 2. Annotation example from one of the specialists.

Therefore, we received a set of 120 definitions annotated by at least two experts; they were three in the case where the first two did not agree. This set includes examples for each of the previous cases. Analysing it, we managed to identify a unique aspect: even if all the definitions follow a certain number of patterns, each of them is almost an individual combination of several regular expressions. Our focus was to identify three meaning relations: synonymy, hypernymy, and meronymy.

We observed and applied the following rules, continuously checking the results after applying, and improved our program with the new ones.

⁵ Features of a particular meaning.

4.1 Regular Expressions Represented by Word Structures

Most of the regular expressions of this type were used to mark the hyperonymy, expressed by sentences starting with:

- “Numele/Numele dat/unui... / The name/Name given/of...” (e.g. “ȘĂMEN = Numele unui # joc # de cărți / = The name of a card #game#”),
- “Varietate a/de... / Variety of...” (e.g. “ȘVÁİȚER = Varietate de # brânză # fină... / = Variety of fine #cheese#...”),
- “Diminutiv/Augmentativ... / Diminutive/Augmentative...” (e.g. “ȘEPCÚȚĂ = Diminutiv al lui # ș a p c ă (2) . # / LITTLE CAP = Diminutive of # c a p (2) . #”),
- “Epitet... / Epithet...” (e.g. “ȘACÁL = *Epitet* pentru o #persoană# *rapace*, *josnică* / = *Epithet* for a *greedy*, *vile* #person#”).

Moreover, we found patterns for meronymy. The first one identified and applied was represented by the definitions that start with “Parte a... or Partea... / Part of...” (e.g. ȘA = Partea %spinării% ... / = Part of %the spinal%...). The other regular expressions for meronymy were not introduced before calculating the statistical results from the following section, but they are mentioned.

Furthermore, we observed that, most usually, at the end of definitions we have synonyms, given by different variants of the searched word. Our program, that annotates based on the token position, will identify the synonyms by words inside the brackets such as:

- “rar / rare” (e.g. “ȘERBEGÍU = # Persoană # care prepara și vindea șerbet ; (învechit , rar) \$ șerbetar \$. / = # Person # that prepared and sold sherbet; (outdated, rare) \$ sherbetist \$”),
- “familiar / familiar” (e.g. “ȘTRENGĂRÍȚĂ = # Fată # sau femeie tânără vioaie, zglobie, ghidușă; (familiar) \$ ștregăroaică \$. / = zestful, playful girl or woman; (familiar) \$ ștregăroaică \$”),
- “regional / regional” (e.g. “ȘELÁR = # Persoană # care ...; (regional) \$ șoronar \$. / SADDLER = # Person # that ...; (regional) \$ șoronar \$”),
- “învechit / outdated” (e.g. “ȘERBEGÍU = # Persoană # care prepara și vindea șerbet ; (învechit , rar) \$ șerbetar \$. / = # Person # that prepared and sold sherbet; (outdated, rare) \$ sherbetist \$”).

It is important to mention that we did not find word structures to represent regular expressions for synonyms; they are identified based on the following two types of patterns.

4.2 Regular Expressions Represented by Abbreviations

We were able to identify synonyms marked by a certain abbreviation (“V.”, “v.”, “Cf.”) that suggests the transfer of meaning between two words. Another formal aspect that interests us in these cases is the way of writing the word, with spaces between the letters.

For example, “ȘÍPCĂ = # Bucată # de lemn subțire și lungă , de obicei de formă paralelipipedică , folosită mai ales în lucrări de tâmplărie sau de dulgherie . V. \$ l e a ț \$ / SHINGLE = A piece of thin and long wood, usually parallelepiped in shape, used especially in carpentry or joinery works. V. \$ l e a ț \$”.

4.3 Regular Expressions Represented by Punctuations & Position

Firstly, we have to mention here the case where the definition is represented by a set of two or more synonyms separated by:

- commas (e.g., for a set of two synonyms, “ȘAFLÍC = \$Căldare\$, \$găleată\$. / = \$scauldron\$, \$bucket\$”),
- semicolon (e.g. “ȘAIT = \$ Canea \$; \$ cep \$ / = \$tap\$; \$bung\$”).

For the situations when this type of sets appears in the middle of definition, the synonyms will be separated of other parts of definition by a semicolon (e.g. “ȘELĂR = # Persoană # care ...; (regional) \$ șoronar \$. / SADDLER = # Person # that ...; (regional) \$ șoronar \$).

Thus, most of the definitions are not based only on the mentioned regular expressions; they are expressed by a sentence that details the described word. This type of defining represents a regular expression itself for nouns, due to the first word being a noun and hypernym at the same time (e.g. “ȘENÁL = # Curs # de apă / = water#course#”).

5 Results & Observations

We filtered the definitions of all the synonyms inside the mentioned volume of DLR and applied the mentioned regular expressions. As a result, we annotated a collection of 3080 definitions, with a total of 3312 elements annotated by the machine; the resulted dataset can be accessed here⁶.

The precision of our algorithm, calculated on the mentioned data, and based on the human validation, is 88,5%; simultaneously, the recall is equal to 82,9%.

⁶ <https://drive.google.com/file/d/1vWN1AY1bTo1FQ3grN7mP6-swdC91SnkF/view>

Table 1. Statistics.

Definitions	Precision	Recall
3080	88,5%	82,9%

The appearance of exceptions and particular cases of the rules are not surprising, given the fact that the three types of patterns combine in different ways; the results of precision and recall statistically confirm it. Observing the resulting dataset, we were capable of justifying the two results and identifying improvement methods for the definitions. Both the recall and precision results relate to the method we proposed, considering that we decided to not preprocess the definitions (excepting tokenization) and focus on the formal appearance.

First of all, the precision result is given either by indexation problems, or by pattern mistakes or particularities. The indexation issues were influenced by different words inserted in the regular expressions: e.g. instead of “Epitet pentru... / Epithet for...”, we found “Epitet la adresa... / Epithet addressed to...”.

At the same time, we have to ensure the correct function of the regular expressions in the following cases: words separated by hyphen and considered two tokens, as in “Varie- tate scumpă de damasc. / Expensive varie- ty of damask.”, set of two hypernyms separated by comma or “sau / or”, marked so far as synonyms.

Regarding the recall result, we observed that the regular expression with abbreviations and words separated in letters is not working accordingly over the entire definition, and so the appearance of the set of synonyms. Furthermore, we identified more regular expressions: words such as “familiar / familiar” at the beginning of the definition, phrases that start with “Specie de... / Species of...”, “Ansamblu de... / Ensemble of...”. Beside these, we found different patterns for meronymy at the beginning of the definition: “Prelungire a... / Extension of...”, “Bucată/Bucăți de... / Piece/Pieces of...”.

We plan to implement all of these observations and improve our precision and recall results as much as possible.

6 Future Work

As future work, after improving the mentioned results, we will focus on:

- (if necessary) gain as many human annotations as possible to train the machine in order to maximise the recall;
- apply the same methods for the definitions of other parts of speech;
- extend our study to other volumes of the dictionaries, observe the results with the regular expression already identified, and maximise them as well;
- develop a suitable graphical representation of all the meaning relations.

We believe that, by fulfilling the directions from before, we will make the Romanian lexical material more accessible to both consult in daily activity and use with professional purposes.

7 Conclusions

Based on our results so far, we can conclude that the given method helps both linguists and programmers, by developing the Romanian material used in NLP and give clarity regarding different linguistics aspects.

Through the lens of Romanian lexical resources in electronic format, we can recognize the value brought by approaching the dictionaries; of course, having in mind the right linguistics approach. By taking this path, we can improve the existing datasets, and create an accessible, user-friendly platform for the interested large public.

References

- Dicționarul limbii române. (DLR).* Serie nouă. 1978. [Redactori responsabili: acad. Iorgu Iordan, acad. Alexandru Graur și acad. Ion Coteanu. Din anul 2000, redactori responsabili: acad. Marius Sala și acad. Gheorghe Mihăilă]. Tomul XI. Partea 1. Litera Ș.
- eDTLR. 2023. <https://gitlab.com/soliorom-cflr/data/edtlr/-/tree/main> (limited access).
- Bahnaru, V. 2009. *Elemente de semasiologie română*, Editura Știința, pages 9-44.
- Dascălu, M. and Gîfu, D. 2015. Evaluating the Complexity of Online Romanian Press. *Proceedings of the 11th International Conference "Linguistic Resources and Tools for Processing of the Romanian language"*, D. Gîfu, D. Trandabăț, D. Cristea, D. Tufiş (eds.), "Alexandru Ioan Cuza" University Publishing House, Iași, pages 149-162.
- Garcia, M. 2021. Exploring the Representation of Word Meanings in Context: A Case Study on Homonymy and Synonymy, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625-3640, Online, Association for Computational Linguistics; <https://aclanthology.org/2021.acl-long.281/>.
- Hauer, B., Bao, H., Mallik, A., and Kondrak, G. 2021. UAlberta at SemEval-2021 Task 2: Determining Sense Synonymy via Translations, *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 763-770, Online, Association for Computational Linguistics; <https://aclanthology.org/2021.semeval-1.101/>.
- Khan, F., Chiarcos, C., Declerck, T., Gîfu, D., González-Blanco García, E., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J., Pagé-Perron, É., Passarotti, M., Ros, S., Truică, C.-O. 2022. When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Open Data. *Semantic Web Journal*, Tracking #: 2859-4073, 13(6):987-1050.

- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. 1993. *Introduction to WordNet: An On-line Lexical Database*. August 2023. <https://wordnetcode.princeton.edu/5papers.pdf>.
- Eventhough, N. 2009. Sense Relations in the Treatment of Meaning in Isichazamazwi SesiNdebele, *Lex*, 19(1):1.
- Peritexts for DLR/DA volumes. Last accessed November 2023. <https://dlr1.solirom.ro/>.
- Popović, A. H. 2023. Synonymy from a Prototype Theory Perspective and its Symbiosis with Polysemy: Towards a New Dictionary of Synonyms. *Lex*, 33(1): 1-23, April 2023.
- Alamillo, A. R., Moreno, D. T., González, E. M., Acosta, M. T., Taroni, A. and Valadez, J. H. 2023. The Analysis of Synonymy and Antonymy in Discourse Relations: An Interpretable Modeling Approach. *Computational Linguistics*, 49(2):429-464.

CHAPTER 6

ROMANIAN SPEECH

MAKING LINGUISTIC RESOURCES ACCESSIBLE. THE AUDIO ARCHIVE OF THE NEW ROMANIAN LINGUISTIC ATLAS BY REGIONS. MOLDOVA AND BUKOVINA

FLORIN-TEODOR OLARIU¹, ALEXANDRU LAURENȚIU COHAL¹,
LUMINIȚA BOTOȘINEANU¹, VERONICA OLARIU¹,
RAMONA LUCA², SILVIU-IOAN BEJINARIU^{2,3}

¹ “Alexandru Philippide” Institute of Romanian Philology,
Romanian Academy Iași Branch, T. Codrescu Street, 2, Iași, Romania
{olariuft, veronica_olariu}@yahoo.com,
{alexandru.laurentiu.cohal, lumi.botosineanu}@gmail.com

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

³ “Iorgu Iordan – Alexandru Rosetti” Institute of Linguistics,
Romanian Academy, 13 Septembrie Way, 13, Bucharest, Romania

Abstract

In this paper we present the general framework of a new project dedicated to the computerization of Romanian geolinguistics, entitled *The New Romanian Linguistic Atlas by Regions. Moldova and Bukovina – linguistic resources in electronic format (eNALR-MB)*. The project will be a complex and interdisciplinary one as a result of the collaboration between linguists and computer scientists from the academic centre of Iași and will aim to make accessible online the existing linguistic resources within the Department of Dialectology and Sociolinguistics from the “Alexandru Philippide” Institute of Romanian Philology from Iași, including the audio archive, the volumes of atlases and dialectal texts published so far, as well as various elements characteristic of dialectal research as: photos, survey sheets, marginal notes on questionnaires, etc. In this paper, the focus is on the audio component of the *NALR-MB* archive, detailing the manner in which it will be structured and made public through a dedicated digital platform created for this purpose.

1 Introduction

Within the design and planning processes of the project entitled *The New Romanian Linguistic Atlas by Regions. Moldova and Bukovina – linguistic resources in electronic format (eNALR-MB)* two components played a major role. First of all, it is about the external context of profile research (the philological one, in particular), characterized, among other things, by the need to preserve, digitize and promote the intangible cultural heritage, this direction of research currently representing one of the imperatives major aspects of international research policies and strategies. Examples such as: *Europe's digital cultural heritage*¹, *World Digital Library*², *Gallica Digital Library*³ or *SONIC* database of the Library of Congress⁴ (about 2.5 million audio recordings of which 350.000 are bibliographic records, Fujinaga and Forscher, 2008) illustrate very well this trend, materialized mainly since the 2000s, when the digitization of cultural resources experienced an exponential development in more and more countries. In their quality as a fundamental component of the intangible cultural heritage, the linguistic resources (dictionaries, atlases, old texts, etc.), have experienced, in turn, an extensive process of digitization and computer processing (Cristea *et al.*, 2016; Gifu *et al.*, 2015), which led to the creation including a new field of research, called *Digital Humanities*.

In terms of dialectal resources, the classic printed format of atlases is the most frequently used for all languages. Web versions accompanied by written documents audio recordings have been developed in the last decade: *Atlas Lingüístico (y Etnográfico) de Castilla – La Mancha*⁵ (ALECMAN), *Linguistic Atlas of Dolomitic Ladinian and Neighbouring Dialects*⁶ (ALD) and *Atlas Lingüístico de la Península Ibérica*⁷ (ALPI). One of the most complete achievements is the *Audiovisual Linguistic Atlas of the Franco-Provençal Dialects of French-speaking Valais (L'Atlas linguistique audiovisuel des dialectes francoprovençaux du Valais romand)*⁸ (ALAVAL) developed at the University of Neuchâtel. Its main window contains four interconnected areas: the map, table with answers and comments, video-player and the list of title words.

¹ www.europeana.eu

² www.wdl.org

³ <https://gallica.bnf.fr/>

⁴ <https://www.loc.gov/rr/record/research.html>

⁵ <https://alecman.web.uah.es/contenido/Accesso%20a%20mapas.htm>

⁶ www.ald.gwi.uni-muenchen.de/en/?db=ald2

⁷ <http://alpi.csic.es/>

⁸ <http://alaval.unine.ch/>

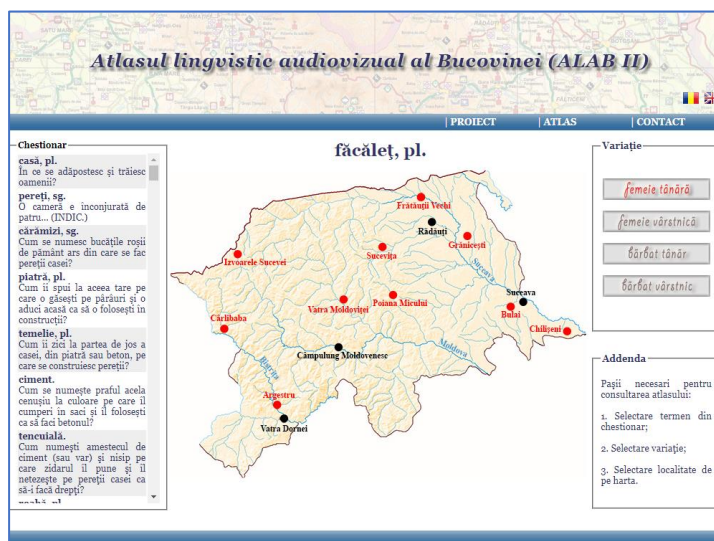


Figure 1. Main window of the Audiovisual Linguistic Atlas of Bukovina (www.philippide.ro/alab/)

The second major impulse in the configuration of the *eNALR-MB* project was the solid joint experience of the linguists and computer scientists from the Iași Branch of the Romanian Academy in the direction of digitizing the process of publishing dialectology and geolinguistics works. This collaboration started at the beginning of the 2000s and was realized, first, in the creation of the first computer application in the field of the Romanian language used for editing linguistic atlases and volumes of dialect texts: *ALRMaps*, with the help of which three volumes of *NALR-MB* have been published (Botoșineanu *et al.*, 2022). The second important moment in the development of this inter-disciplinary collaboration was represented by the design and development of the first linguistic atlas in audio-visual format dedicated to some dialects of the Romanian language (and one of the first of its kind at the international level): Audiovisual Linguistic Atlas of Bukovina (ALAB) whose web page is shown in Figure 1 (Olariu *et al.*, 2016). Another Romanian approach is that of the Linguistic Atlas of Romanian Dialects from North and South of the Danube, which is still being implemented in “Iorgu Iordan – Alexandru Rosetti” Institute of Linguistics of the Romanian Academy. Apart from the printed form of the atlas, an online version with multimedia facilities is also available. Pictures, audio, and video recordings as well as documents related to the interviews and localities where they took place are available through the web interface presented in Figure 2 (Bejinariu *et al.*, 2023; Iftene *et al.*, 2023).

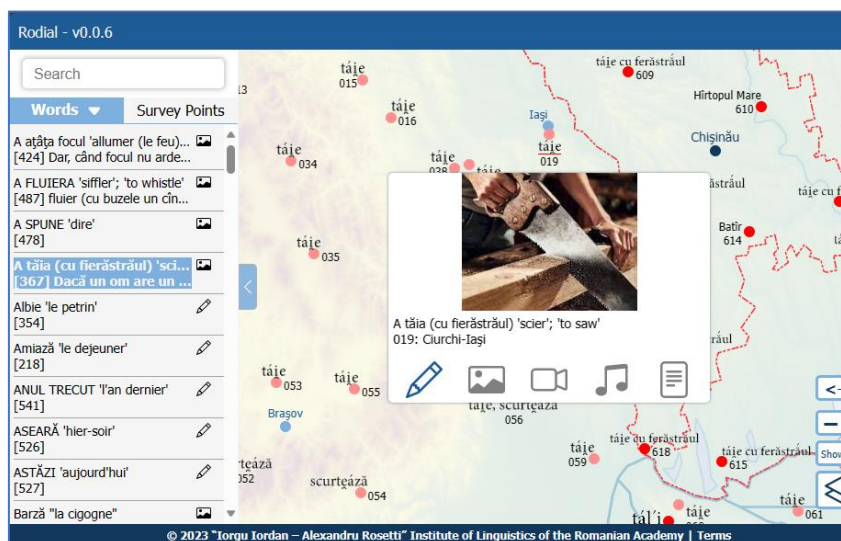


Figure 2. Digital map in the on-line version of the Linguistic Atlas of Romanian Dialects from North and South of the Danube.

These notable results (most of them being national premieres) obtained in the last 20 years (Bejinariu *et al.*, 2021; Botoșineanu *et al.*, 2010) have turned into a consistent motivation towards the further development of this innovative spirit that has characterized recent research in the field of geo-linguistics in Iasi, by setting up a new project that will highlight the entire research experience accumulated by the members of the Department of Dialectology and Sociolinguistics in the last 6 decades, when they started designing and, subsequently, carrying out field surveys for *NALR-MB*. The project is described in the next sections of the paper. The second section describes the linguistic resources which will be published. The structure of the system and some implementation details are presented in the third section, and the last one concludes the paper.

2 The Linguistic Materials Available in the *NALR-MB*

In order to reproduce in as much detail as possible the stages and the results obtained in the about 60 years of activity, the structure of the *eNALR-MB* platform was designed on three major components:

- *published works* – five volumes of *NALR-MB* and three volumes with dialectal texts will be scanned and uploaded to the platform. In the dialectal archive thus constituted, it will be possible to carry out automatic search operations according to several criteria;

- *sound map* of the Moldavian dialects on the right side of the Prut River. It will be an online map of Moldova and Bukovina that will contain points from the *NALR-MB* network in which the users will have the possibility to listen speech samples and complete recordings from the archive;
- *varia* - snapshots from the history of the research activities, such as photographs during the investigations, notes on the side of the questionnaires, explanatory notes and sketches of the investigators from the investigation notebooks, elements specific to the different stages of the to the digitization process of editing the atlas volumes, etc.

Through *eNALR-MB*, the channels for disseminating information multiply, for the benefit of the target audience, who will thus come into contact not only with the written reflection of the Moldovan-type linguistic varieties, but also with their concrete phonetic expression, in a position to overcome the inherent limitations of printed volumes, despite the assumed accuracy of phonetic transcription.

2.1 Published Linguistic Atlases and Dialectal Texts

The design of the volumes of texts (TD-MB I/1, 1993; I/2, 1995; II/1, 2002, Figure 3) as an integrant part of the Romanian linguistic atlases, explicitly and programmatically assumed by the *NALR-MB* authors, as an aspect through which Romanian dialectology is positively individualized in the whole of Romance linguistic geography, ensures continuity with the atlas, with all the positive consequences arising from the unity of vision of the authors-investigators, from the common network of points or from the possibilities of completion, clarification and contextualization offered by the consistent inclusion of the informants used for the questionnaire on the list of subjects from which were also recorded texts.

In the context of this solid tradition, embodied not only in the two volumes (in three toms) of *TD-MB* published so far, but including all the other profile works that accompany and complement the Romanian regional atlases, the need to connect to the problematics and methodology of these previous achievements is outlined. Through our approach, making available to the interested public the original dialect material stored in the *NALR-MB* sound archive, an objective achieved so far by the elaboration of dialect anthologies in classic format, is not only enriched, proposing, as an alternative solution, the online publication of some dialectal speech samples, but also continues, integrating the working principles and methodological solutions tested by our colleagues from the generation that, unfortunately, is gone now.

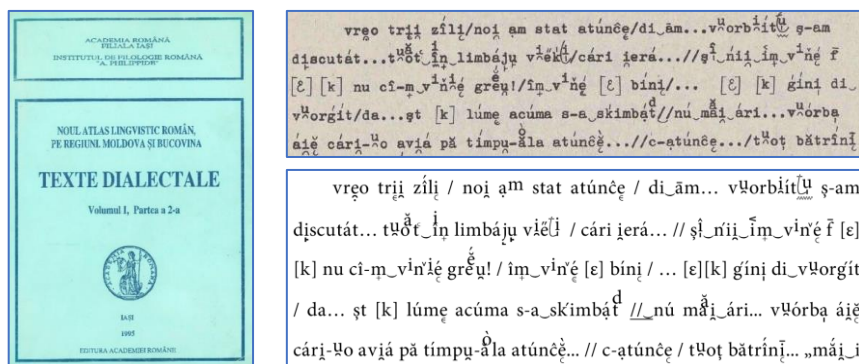


Figure 3. Cover and dialect text sample from previously published TD-MB I/2; the same text edited using *AlrWord* add-in (Bejinariu *et al.*, 2022).

For the *TD-MB* elaboration, between 8 and 30 informants were registered in each survey point, each of them with up to 20 texts, thus resulting a total of 50 to 140 texts for each locality. For this reason, for publication, a selection of this immense material was self-imposed. In addition to the speaker's representativeness, narrative talent and amiability, the selection criteria also took into consideration the intention to cover, as good as possible, all the categories of targeted informants and all the types of these materials. From this perspective, a landmark for us was the objective of the authors of *TD-MB* previously published to retain for each survey point a minimum of one to two texts for one to three informants from each age group. For the group of informants, in relation to the diagenational and diasexual stratification, a simplification was made in *eNALR-MB*; it will consider 6 categories, grouped on three generational strata: children, young adults and elderly adults, combined with the parameter represented by the biological gender.

The limitations of space, time and human resources also suggested an experimental restriction of the *NALR-MB* survey point's network to those which are shared with the *ALR II* network. Through this methodological option we intend to provide in Table 1 eloquent material for diachronic projections.

2.2 Audio-archive of *NALR-MB*

The sound map is the central point in the economy of the entire *eNALR-MB* project, which is why in the present paper is focused on detailing the way in which the audio archive will be structured and made public through a digital platform specially configured by IT specialists.

The archive contains over 600 tapes with more than 1000 hours of recordings from the 210 localities of the *NALR-MB* network. These tapes were recorded

during the dialectal investigations from 1960-1975 and are available at the Department of Dialectology and Sociolinguistics of the Institute of Romanian Philology. The preparation of the recordings in order to include them in the audio

Table 1. Survey points in eNALR-MB.

Nr.	Village	Commune	County	Cartographic number in	
				<i>NALR-MB</i>	<i>ALR II</i>
1	Brodina	Brodina	Suceava	465	366
2	Ciocănești	Iacobeni	Suceava	469	365
3	Marginea	Marginea	Suceava	478	386
4	Cristești	Cristești	Botoșani	515	414
5	Larga Jijia	Movileni	Iași	530	520
6	Coropcenii	Ciorțești	Iași	536	514
7	Mircești	Mircești	Iași	550	537
8	Pipirig	Pipirig	Neamț	570	551
9	Măgura	Măgura	Bacău	592	531
10	Furceni Vechi	Cosmești	Galați	654	605

archive consisted of going through the following stages (Cohal *et al.*, 2022):

- the recordings made in the 10 survey points considered in the project were analyzed in order to select those of good quality, after which they were digitalized in Waveform Audio File Format (.wav) resulting in a recorded material of about 300 minutes;
- improving the quality of audio files in the middle frequency segment (that of the human voice) using the *Adobe Audition* application;
- in the audio files, the segments related to the topics of interest decided by the linguistic researchers were tagged (Figure 4).

In addition to the basic operations of segmenting audio files, increasing or decreasing volume, speeding up or decelerating the time of the original recordings, some plugins in *Adobe Audition 2022* and *Izotope RX 10* have allowed a better quality of our audio tracks, and among these we recall: *Fast compressor* (to smooth out volume spikes); *FabFilter Pro Q3* (equalizer, to remove low frequencies and annoying frequencies); *Izotope RX Denoise* (to intelligently act on noise on high frequencies); *Oeksound Soothe2* (to attenuate resonances); *Slow compressor* (to render more natural vocal frequencies) and *Limitter* (to establish a consistent output volume across tracks).

Thus, from the complete recordings of the surveys from the 60s, they were obtained, for the *eNALR-MB*, a number of about 100 selections, each with a

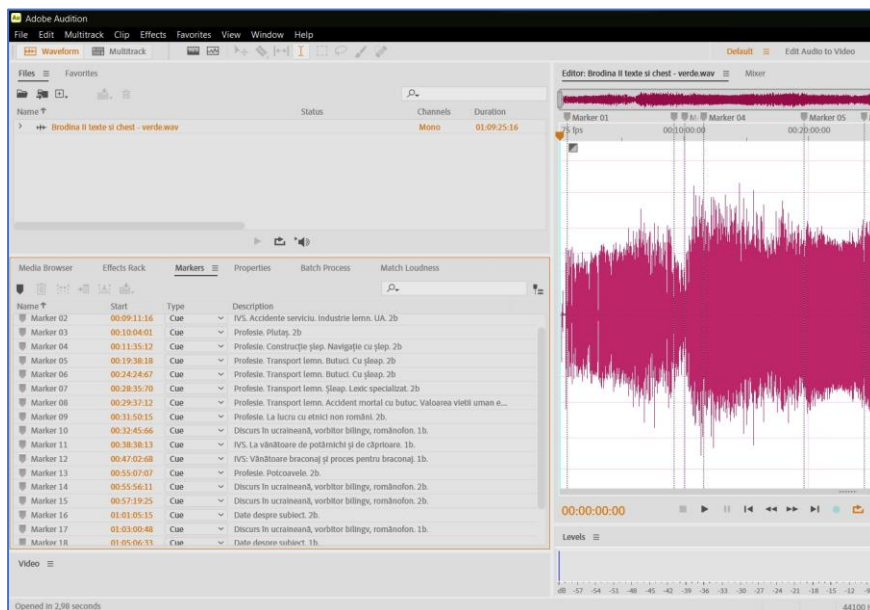


Figure 4. Audio files segmentation
(screenshot from Adobe Audition application).

duration of about 2 minutes. Each audio file then received a coded name based on the agreed selection criteria, a name that contains the essential elements of its description: cartographic number, subject of the recording, location of the action, diagenational and diasexual variation, name of the investigator. Thus, in the sound map, visitors to the site will be able to access good and very good quality audio files from Moldova and Bucovina in the 1960s-1970s that will be easy to individualize, access and possibly download for further linguistic analysis.

3 Electronic Version of *NALR-MB*

The audio archive is mainly a collection of recordings that preserves the general characteristics of a database: consistency, atomicity, and data isolation, which allows the sharing of the included information but whose security and integrity must be ensured through the use of a system (Löhr and Rakow, 1995). In particular, since it is an audio collection, it is necessary to store information related to (Lutfi *et al.*, 2004): audio file format (encoding type, duration); measurable characteristics of the signal (quality); the subject or theme of the recording; the context in which they were made (type and location); annotations (e.g. creators, topics, etc.).

The audio archive is available for consultation on the Web through the pages designed for this purpose. The actual access to the recordings is made directly by browsing the database or by selecting the survey point on the digital map.

3.1 Data Model

The information in the *eNALR-MB* database is organized in nine tables (Figure 5) that contains not only the audio collection but also data related to the linguistic atlases and dialectal text.

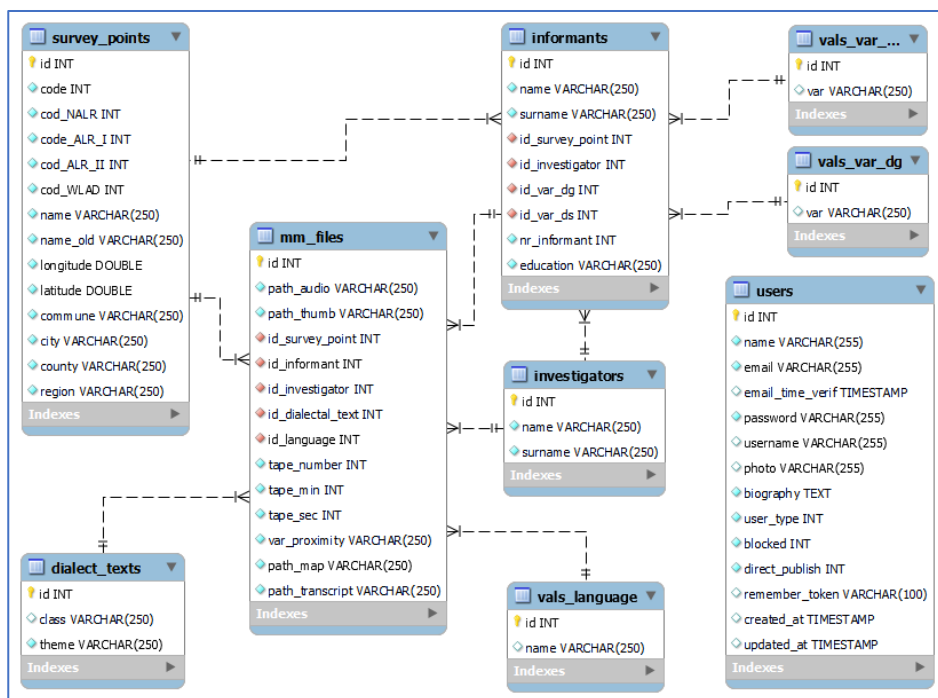


Figure 5. Structure of the eNALR-MB database

The specific content of the main tables is, as follows:

- **Survey points** table contains information on the locations where the dialect surveys were conducted. In addition to the NALR, ALR I, ALR II, WLAD codes the table also contains information about their geographical coordinates, the current and old name, commune, city, county and the region of which it is a part.
- **Investigators** table contains minimal data, such as first and last name of the researchers who conducted the surveys.

- **Informants** table contains information about the subjects of the *NALR_MB* surveys. In addition to the name, surname and education level of the informant, other information such as the point where the survey was carried out, investigator, the diagenational and diasexual variations is also included.
- **Mm_files** is the main table in the database that contains the audio recording file information (name and path) as well as the links to related information (survey point, investigator, informant, etc.). The table contains more records for the same survey: the audio or video recording, images and the document that contain the phonetic transcription of the interview.

The database is completed with tables in which the dialectal theme (*dialect_texts*), diagenational variation (*vals_var_dg*), diasexual variation (*vals_var_ds*) and languages (*vals_language*) are encoded as well as the user authentication management table (*users*).

3.2 Digital Map of Moldova and Bukovina

For an easy and intuitive access to the audio archive the survey points can be selected on the digital map of Moldova and Bukovina (Figure 6) which is designed on the GIS model (Bejinariu *et al.*, 2016). The map is prepared using the free open-source QGIS geographic information system package.

It is a simplified version of the map used for *NALR-MB* atlases publication. The ESRI shape file format is used to store the map's content that is organized into vector layers based on the geographic features they represent: map contour (used for designing purposes), external borders, internal (regional) borders, major water courses, name of water courses, cities and survey points. The relief is represented on the map using a single raster layer: background. This list contains the layers that are used to draw the map of Moldova and Bukovina (Figure 6). However, it contains some more hidden layers that are required if the application needs to be adapted for other regions, e.g. major lakes, maritime borders or capitals.

3.3 Data Management Interface

The *eNALR-MB* database is implemented using the MySQL database management system. The following tools and technologies are also used for the development of the web interface: PHP (is a server-side scripting language designed specifically for creating dynamic web applications, HTML and CSS), Laravel Framework (open-source PHP framework designed for the rapid, simple, and organized development of web applications using the PHP language), Node.js (open-source server-side runtime environment for JavaScript) and Bootstrap (open-source CSS framework that provides design templates for web interfaces based on CSS and JavaScript).

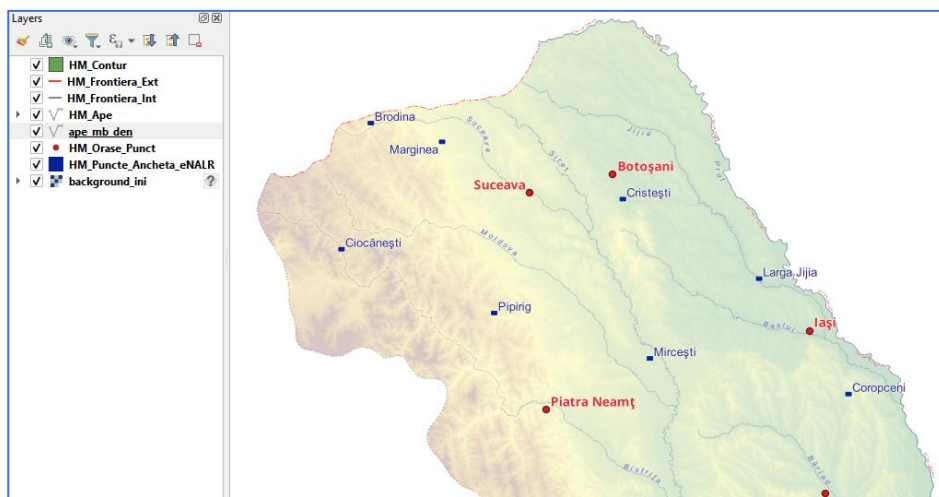


Figure 6. Digital map of *eNALR-MB* (screenshot from QGIS application).

The access to the data management web interface is secured through user and password authentication.

Figure 7. Interface for CRUD operations on survey points.

The information about linguistic resources from *NALR-MB* are registered in the *eNALR-MB* database through a web interface using data entry forms. The top header of the interface contains the following links (Figure 7): *Puncte de anchetă* (survey points), *Teme etnotexte* (themes of ethnotexts), *Anchete dialectale* (dialectal surveys), *Informatori* (informants), *Anchetatori* (investigators), *Variații* (variation) and *Limba vorbită* (spoken language). Each link opens a web page with

several tabs for CRUD (create, read, update, and delete) operations of data. In Figure 7, the page is accessed by clicking the *Survey Points* link and contains three tabs, two for adding data about the survey points into *eNALR-MB* database using data entry forms and the third for reading, updating, or deleting data.

The link *Informants* opens the interface for CRUD operation of the subjects who speak on the audio files (Figure 8).

The form is titled 'Aadaugă informatori' and includes a 'Lista informatori' link. It contains several input fields and dropdown menus for adding informant information. The fields are arranged in two columns. The first column includes: 'Aadaugă nume informator' (text input), 'Aadaugă prenume informator' (text input), 'Selectează anchetator' (dropdown), and 'Selectează punct de anchetă' (dropdown). The second column includes: 'Selectează variația diasexuală' (dropdown with 'Bărbătesc' selected), 'Selectează variația diagenerațională' (dropdown with 'Copil' selected), 'Aadaugă număr informator' (text input), and 'Aadaugă educația informatorului' (text input with placeholder 'Câte clase școlare are informatorul.'). A 'Salvează' button is located at the bottom left.

Figure 8. Interface for CRUD operations about subjects who speak on audio files.

The editing form of the subject's information contains fields for adding data about name, surname, age and gender variation of the subject, its level of education and the survey point. To express of the diagenerational and diasexual variation, the possible values for age are children / young adults / elderly adults and male / female respectively are shown in Figure 9.

The table is titled 'Variații / Limba vorbită' and is divided into three sections: 'VARIATIE DIAGENERAȚIONALĂ', 'VARIATIE DIASEXUALĂ', and 'LIMBA VORBITĂ'. Each section has a table with columns for ID, variation name, and actions (Modifică | Șterge).

#	VARIATIE DIAGENERAȚIONALĂ	UPDATE / DELETE	#	VARIATIE DIASEXUALĂ	UPDATE / DELETE	#	LIMBA VORBITĂ	UPDATE / DELETE
1	Copil	Modifică Șterge	1	Bărbătesc	Modifică Șterge	1	română	Modifică Șterge
2	Tânăr	Modifică Șterge	2	Femelesc	Modifică Șterge	2	huțulă	Modifică Șterge
3	Bătrân	Modifică Șterge						

Figure 9. Interface for CRUD operations for age and gender variations and spoken language.

The topics of the audio files are accessed with the link *Themes of ethnotexts* in Figure 10 and also, CRUD operations are available for these data. For the audio recordings chosen to be added in the audio archive, the possible values for the themes are rafting, hovering accident, hunting, how to make a wheel, how to make a horseshoe, how to make a wooden house, carpentry - the roof of the house, the meeting of relatives from the USSR, how to cut a pig, and bathing accident. This list can be extended if required.

Aadaugă temă etnotext		Lista temelor etnotextelor	
#	TEME ETNOTEXTE	UPDATE / DELETE	
1	Plutărit	Modifică	Șterge
2	Accident la plutărie	Modifică	Șterge
3	Vânătoare	Modifică	Șterge
4	Cum se face roata	Modifică	Șterge
5	Cum se face potcoava	Modifică	Șterge
6	Dulgherie acoperișul casei	Modifică	Șterge
7	Cum se face casa de lemn	Modifică	Șterge
8	Întâlnirea rudelor în URSS	Modifică	Șterge
9	Cum se taie porcul	Modifică	Șterge
10	Accident la scăldat	Modifică	Șterge

Figure 10. Examples of topics of the audio files.

For entering data into *mm_files* which is the main table of the database, the interface shown in Figure 11 is used. The multimedia file names and all associated information can be selected using the drop-box controls, as they display data already entered into the database, such as survey points, topics, spoken language, the names of the subjects and the persons who records the audio files. The proximity variable has three options: *în gospodărie* (household chores), *în comunitate* (in the community) and *în afara comunității* (outside the community). Also, the form contains controls in which the audio files duration in minutes and seconds is specified.

After all the necessary information is entered into the database, they are available for consultation in tabular format. The interface that allows selection directly from the map is still being implemented.

Figure 11. Interface for CRUD operations on audio files data.

4 Conclusions

The use of information technologies is the optimal solution for safeguarding and promoting intangible cultural heritage. Dialectology and geolinguistics, traditionally focused on the ethnolinguistic component of this heritage, make full use of this new opportunity to develop their own research fields. The *eNALR-MB* project tries to respond to this desideratum of the digitization of ethnolinguistic resources and to make available to both specialists and the interested public a large volume of audio information or the printed format existing within the Department of Dialectology and Sociolinguistics from Iasi, representative of the linguistic specificity of the area between the Carpathians and the Prut River. Having as its main point of interest the sound map of the Moldovan dialects on the right side of the Prut, the *eNALR-MB* platform will finally include the volumes of atlases and dialect texts that have been published so far, thus becoming the main digital resource capable of illustrating a detailed ethnolinguistic profile of Moldova and Bukovina. In addition to making the audio archive accessible to the public, the software solution proposed in this project is a research tool for dialectologists whose usefulness and value will be demonstrated when the audio archive will be completed for all survey points not only in the Moldova area but in the whole Romanian linguistic area.

Acknowledgements

The results presented in this paper were obtained within the framework of the fundamental research projects of the Romanian Academy. A preliminary version of this research was presented in (Cohal *et al.*, 2022; Olariu *et al.*, 2022) at the international workshop *Romanian Linguistic Geography in the Era of Digitalization*, October 21-22, 2022, Iasi, <https://sites.google.com/view/geolinguistics-2022>.

References

- Bejinariu, S.-I., Luca, R., and Olariu, F.-T. 2016. A GIS Based Approach for Information Management in Geolinguistics. *Memoirs of the Scientific Sections of the Romanian Academy*, Tome XXXIX, Computer Science, pages 37-45.
- Bejinariu, S.-I., Iftene, F., Nevaci, M., and Floarea, C. I. 2021. Preservation of Romanian Linguistic Heritage. Framework for Dialectal Data Management. In P. Rebeja, M. Onofrei, D. Cristea, D. Tufiş (eds.), *Proceedings of the 16th edition of The International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR 2021*, 13-14 December 2021, Iaşi, pages 75-92.
- Bejinariu, S.-I., Luca, R., Apopei, V., and Iftene, F. 2022. Dialect Text Editing Using the Phonetic Transcription of the Romanian Language (Original title in Romanian: Editarea textelor dialectale folosind transcrierea fonetică specifică limbii române). In *Anuar de lingvistică și istorie literară*, Tom LXII. IN MEMORIAM STELIAN DUMISTRĂCEL, pages 37-53.
- Bejinariu, S.-I., Nevaci, M., Saramandu, N., and Iftene, F. 2023. Modern Techniques for Linguistic Atlases Digitalization. Applications in ALDRo - Atlas of Romanian Dialects from the North and South of the Danube. *Colloquium Romanian dialects and Romance dialectology*. 9 June 2023, Zürich, Switzerland.
- Botosineanu, L., Olariu, F.-T., and Bejinariu, S.-I. 2010. A Computerization Project in Romanian Linguistic Cartography: The New Romanian Linguistic Atlas by Regions. Moldova and Bukovina in Electronic Format, (e-NALR) - Achievements and Perspectives (Original title in French: Un projet d'informatisation dans la cartographie linguistique roumaine: Noul Atlas lingvistic român, pe regiuni. Moldova și Bucovina en format électronique, (e-NALR) – réalisations et perspectives). *XXVIth International Congress of Linguistics and Romance Philology*, 26 CILFR València 2010, September 6-11, 2010, pages 456-457.
- Cohal, A.-L., Botoșineanu, L., Luca, R., and Bejinariu, S.-I. 2022. Making Language Resources Accessible through eNALR-MB. Phase 1: audio archive (Original title in Romanian: Accesibilizarea resurselor lingvistice prin eNALR-MB. Faza 1: audioarhiva). *The International Workshop Romanian Linguistic Geography in the Era of Digitalization*, October 21-22, 2022, Iasi (onsite & online).
- Cristea, D., Gifu, D., Niculiță, M., Pistol, I., and Sfirnaciuc, D. 2016. A Mixed Approach in Recognising Geographical Entities in Texts. *Linguistic Linked Open Data*, D. Trandabăț and D. Gifu (eds.), Romania, July 13-25, 2015, Revised Selected Papers, © Springer-Verlag Berlin Heidelberg, pages 49-64.

- Fujinaga, I. and Forscher, S. 2008. Music. In Susan Schreibman, Ray Siemens and John Unsworth (eds.). *A Companion to Digital Humanities*. Wiley-Blackwell.
- Gifu, D., Pistol, I., and Cristea, D. 2015. Corpus of Entities and Semantic Relations with Application in Geographical Domains. *Proceedings of the 11th ConsILR*, D. Gifu et al. (eds.), "Alexandru Ioan Cuza" University Publishing House, Iași, pages 67-78.
- Iftene, F., Saramandu, N., Nevaci, M., and Bejinariu, S.-I. 2023. Web Application for Multimedia Content Management in Online Linguistic Atlases, *The 10th Congress of the International Society for Dialectology and Geolinguistics*, September 4-8, 2023, Bucharest, Romania.
- Löhr, M. and Rakow, T. C. 1995. Audio Support for an Object-Oriented Database-Management System. *Multimedia Systems*, 3:286-297.
- Lutfi, R., Gelgon, M., and Martinez, J. 2004. Structuring and Querying Documents in an Audio Database Management System. *Multimedia Tools and Applications*, 24:105-123.
- Olariu, F.-T., Olariu, V., Bejinariu, S.-I., and Luca, R. 2022. NALR. Moldova and Bucovina – Linguistic Resources in Electronic Format - eNALR-MB (Original title in Romanian: NALR. Moldova și Bucovina – resurse lingvistice în format electronic - eNALR-MB). *The International Workshop Romanian Linguistic Geography in the Era of Digitalization*, October 21-22, 2022, Iasi (onsite & online).
- Olariu, F.-T., Olariu, V., and Luca, R. 2016. Audiovisual Linguistic Atlas of Bucovina (ALAB) - Methodological Benchmarks (Original title in Romanian: Atlasul lingvistic audiovizual al Bucovinei (ALAB) – repere metodologice). In D. Loșonți, V. Vlasin, and N. Mocanu (eds.). *Lucrările celui de-al XVI-lea Simpozion internațional de dialectologie*. Cluj-Napoca, Argonaut & Scriptor, pages 327-329.
- TD-MB. The New Romanian Linguistic Atlas by Regions, Moldova and Bukovina, Dialectal Texts (original in Romanian: Noul Atlas lingvistic român, pe regiuni. Moldova și Bucovina, Texte dialectale). *Publishing House of the Romanian Academy*, Bucharest. Three volumes: Vol. I/1, 1993, Investigation by Stelian Dumistrăcel, published by Doina Hreapcă and Ion-Horia Bîrleanu; Vol. I/2, 1995, Investigation by Stelian Dumistrăcel, published by Doina Hreapcă and Ion-Horia Bîrleanu; Vol. II/1, 2002, Investigation by Ion-Horia Bîrleanu, published by Doina Hreapcă and Ion-Horia Bîrleanu, preface by Stelian Dumistrăcel.

TOWARDS SPEECH-BASED WEB APP DEVELOPMENT

IRINA VASILITA¹, ADRIAN IFTENE¹

¹*“Alexandru Ioan Cuza” University of Iasi, Faculty of Computer Science
{vasilitairina, adiftene}@gmail.com*

Abstract

The development of software applications is a complex and time-consuming process that requires expertise in multiple areas such as programming languages, frameworks, and software architecture. While there are numerous tools available to assist developers in creating software, the process remains challenging, especially for those who lack programming experience. This issue has led to the exploration of new technologies that can help simplify the development process and increase efficiency. Code generation using machine learning has gained significant attention in recent years, due to its potential to automate software development and increase developer productivity. This paper aims to explore the potential of a speech-based code generation system to generate accurate and usable application code directly from the natural language descriptions of the app. Specifically, the research will investigate the feasibility of this approach, examining the technical challenges involved and identifying the most effective techniques and algorithms for generating high-quality code. Additionally, the paper will explore the limitations of this approach and will evaluate its effectiveness in real-world development scenarios.

1 Introduction

In recent years, the use of natural language processing (NLP) and machine learning (ML) has gained popularity in the development of software tools. Many technologies have been developed to simplify the software development life-cycle. However, there is still more to be researched in this area. This paper will focus on leveraging NLP and ML techniques to solve the task of code generation based on spoken descriptions. Based on existing literature, we will start an iterative process to identify the most effective techniques to process the natural language descriptions and the most suitable machine learning models for code generation. In particular, we focus on developing a methodology that will minimize the environmental impact of code generation. By iterating the process,

we want to ensure that we evaluate the found techniques and methodology to ensure that they are efficient and sustainable.

How can we help simplify the development process and increase efficiency? One promising approach is the use of a speech-based code generation system to generate code from spoken descriptions of the app. This approach has several potential benefits. *First*, it can significantly reduce the time required for software development by allowing developers to generate code more quickly and accurately from spoken descriptions. This can be particularly beneficial for complex projects where the development process can be time-consuming and resource-intensive. *Second*, a speech-based code generation system can improve accessibility for developers who have difficulty typing. This includes individuals with physical disabilities or those who are simply more comfortable speaking than typing. By allowing developers to use their voice to generate code, the system can increase productivity and efficiency, while also promoting inclusiveness in the development process. *Third*, accurate code generation may reduce the likelihood of bugs and errors. By allowing developers to describe their apps in natural language, the risk of mistakes and misunderstandings may be reduced. This can ultimately lead to more reliable and efficient software. *Fourth*, the ability to generate code from spoken descriptions may reduce the learning curve for new developers. By having the code automatically generated from natural language descriptions, new developers can more quickly get up to speed with the development process. This can reduce the need for extensive training and support, while also promoting knowledge sharing and collaboration within development teams. Additionally, the generated code can serve as a reference point for new developers to better understand the underlying structure of the software and how it operates. This can lead to a more effective onboarding process for new developers, ultimately resulting in increased productivity and faster time-to-market for software products. *Finally*, a speech-based system has the potential to allow businesses to deliver software faster and at a lower cost. By reducing the time and effort required for software development, businesses can more quickly respond to changing market conditions and customer needs. This can lead to increased competitiveness and improved customer satisfaction.

To ensure timely completion of this research, certain limitations have been set. The scope of this paper will be limited to simple web applications developed using HTML, CSS, and JavaScript, and the generated code will serve only as a starting point for developers to build upon. In terms of evaluation metrics, this work will focus solely on existing metrics and will not introduce any new metrics. Furthermore, due to computing and memory limitations, the dataset used in this research will be relatively small. However, there is a possibility of extending the dataset in the future. The target group for a speech-based code generation system

would primarily be the programmers who face challenges with typing, and those seeking efficient ways to create code for web applications (Khan *et al.*, 2022; Trandabat and Gifu, 2017). However, such a system could also potentially be useful for individuals who are not proficient in programming languages but still need to create simple web applications, such as small business owners or individuals with limited technical expertise. Additionally, the system could be useful for researchers or educators who need to quickly generate code snippets for educational purposes or testing hypotheses (Țucă and Iftene, 2017).

2 Existing Work

2.1 Machine Learning Techniques for Code Generation

Code generation from natural language descriptions has been studied using different techniques, such as neural machine translation, deep learning-based models, and transformer-based models. In the paper (Allamanis *et al.*, 2018), the authors provide an overview of recent advances in using machine learning techniques for code generation and analysis. The authors introduce the concept of “big code” which refers to the vast amount of source code that is available today and explain how machine learning techniques can be applied to this data to improve code generation and analysis. They also discuss various challenges and limitations of using machine learning techniques for code generation and analysis. One of the main challenges is the lack of high-quality labeled data, which is necessary for training machine learning models. Additionally, the complexity and variability of programming languages can make it difficult for models to generalize to new code.

The text-to-code technique involves converting natural language into executable code. This area is closely related to semantic parsing in natural language processing. Several machine learning models have been proposed for text-to-code, including those that generate code in languages used by human software developers, such as Java, Python, and SQL. Applications of text-to-code include converting natural language to Java expressions (Gvero and Kuncak, 2015), shell commands (Lin *et al.*, 2017), simple if-then programs (Quirk *et al.*, 2015), regular expressions (Kushman and Barzilay, 2013), and SQL queries (Zhong *et al.*, 2017), APIs (Gu *et al.*, 2016). Yin and Neubig (Yin and Neubig, 2017) have recently presented a neural architecture for general-purpose code generation. Future research in this area could focus on developing more accurate and efficient machine learning models for big code and naturalness, and on exploring new applications of machine learning in software engineering.

In paper (Sun *et al.*, 2020), the authors propose a novel transformer-based architecture for code generation tasks, which leverages the hierarchical structure

of the code by using an abstract syntax tree (AST) representation. This architecture, called TreeGen in their paper, adopts the Transformer architecture, which is capable of capturing long dependencies. However, their architecture combines the advantages of both tree-based and transformer-based models to improve the performance of code generation tasks. The article (Yin and Neubig, 2017) presents a new approach to code generation using a neural model that takes into account the syntactic structure of programming languages. The authors argue that traditional approaches to code generation do not capture the full range of syntactic structures in programming languages, leading to limited flexibility and accuracy in generating code. The architecture of the model consists of two main components: a neural network encoder and a neural network decoder. The model was trained on different datasets of Python code and natural language descriptions (HEARTHSTONE (HS) dataset (Ling *et al.*, 2016), DJANGO dataset (Oda *et al.*, 2015), IFTTT dataset (Quirk *et al.*, 2015)). The authors evaluated the model on several benchmarks, including generating code from natural language descriptions. The model achieved state-of-the-art results and demonstrated the ability to generate syntactically correct and semantically meaningful code. In paper (Wang *et al.*, 2022), the authors propose the CodeRL framework, which is an approach to code generation tasks that combines pre-trained models and deep reinforcement learning. The model is composed of two main components: a pre-trained language model and a deep reinforcement learning-based policy network. The pre-trained language model used in CodeRL is CodeT5, which is a variant of the T5 transformer model that has been pre-trained on a large corpus of code curated from GitHub, i.e. CodeSearchNet (CSN) corpus (Husain *et al.*, 2019). CodeT5 is capable of generating code for various tasks such as code summarization, code completion, and code translation.

2.2 Code Generation based on Speech

With the increasing prevalence of smartphones and tablets in our daily lives, speech-based inputs have gained significant momentum as a preferred method of human-computer interaction. Voice-based interactions offer a natural and efficient way for users to interact with devices and applications. Voice-based search engines (Aronowitz *et al.*, 2022), such as those found in virtual assistants like Siri, XiaoIce, Google Home, and Cortana, and have revolutionized the way we search for information. By simply speaking our queries, we can obtain instant results and access information with ease. Additionally, voice-based AI assistants and chatbots have become increasingly popular, providing users with personalized and interactive experiences. These intelligent assistants leverage voice commands to perform tasks, answer questions, and assist with various activities. While the field of software development has witnessed significant advancements in the field,

there remains limited research focused on the generation of code using speech. This paper is primarily focusing on capturing natural language descriptions for code generation. However, it is important to acknowledge that research efforts have been dedicated to exploring the use of language-specific voice commands for code generation (Begel, 2005) and for the control of integrated development environments (IDEs) (Maloku and Pllana, 2016; Rosenblatt, 2017).

3 Method

This section describes the technical documentation for the development of our speech-based code generation system. The following questions are answered throughout this section: (1) *How was the dataset constructed, and what steps were taken to ensure its quality and relevance to our research?* (2) *What were the goals and limitations of the methods chosen for our speech-based code generation system, and how were they addressed during development?* (3) *How can the methods used in our system be applied to other datasets or applications beyond our specific research focus?* By providing detailed answers to these questions, this chapter serves as a comprehensive guide to the technical documentation for the development of our speech-based code generation system and may also be useful to researchers and developers working in related fields. Section 3.1 presents how the data was gathered and prepared. Section 3.2 describes the machine learning model used for the code generation task, and Section 3.3 explores the integration of the machine learning model into a web application.

3.1 Code Generation Data

To evaluate the task of code generation, it is important to collect a sufficient amount of raw data and structure it appropriately. In this section, we address the following questions related to data collection and preparation: (1) *What type of data is necessary to perform the code generation task effectively?* (2) *Where can we obtain the required data, and what sources are suitable for this task?* (3) *Should we evaluate the quality of our data to assess its accuracy and usefulness for our research purposes?* (4) *What preprocessing steps are necessary to ensure the data meets the requirements of our study, such as removing irrelevant information or correcting errors?* (5) *What format should the data be in, and how can we describe its features, such as length and complexity?* By addressing these questions, this section provides a clear and comprehensive overview of the data collection and preparation process necessary for effective code generation research.

Data Requirements - *Firstly*, the dataset must consist of a significant number of natural language descriptions that accurately and clearly describe the intended functionality of web applications. These descriptions should be of high quality

and cover a wide range of scenarios to ensure that the code generation system is robust and capable of generalizing to different real-world situations. *Secondly*, the dataset should include corresponding code examples for each natural language description. The code examples must be accurate, functional, and well-structured, closely matching the functionality described in the natural language description. The quality of code examples must also be of high quality, and relevant to the corresponding description.

Data Collection - Based on the requirements outlined in the previous section, we have compiled a dataset consisting of natural language descriptions paired with code snippets. The code snippets were specifically selected based on the descriptions of the web applications provided in the natural language descriptions. The descriptions were selected only in the English language. Given the requirement for high-quality data, it was important to carefully select appropriate sources for data gathering. While GitHub and other similar sources may contain low-quality code and code structure, we were able to gather data from a variety of sources on the internet, including GitHub, personal projects, CodePen, and various tutorials. However, we carefully reviewed and manually selected each data sample to ensure that it met the necessary quality standards.

Data Format - The natural language descriptions in our dataset are presented in plain text format, which is more suitable for a speech-based code generation system than markdown format. Markdown may not be easily processed by certain NLP and machine learning models, making the plain text a more appropriate choice. The plain text format allows for easy processing and understanding by the system. Additionally, presenting the descriptions in plain text makes it easier for other researchers to use the dataset for different tasks, as they can easily extract and manipulate the text data. The descriptions provide clear and accurate explanations of the intended functionality of web applications, allowing the code generation system to learn the relationships between natural language descriptions and code snippets. The code snippets themselves are language-agnostic, meaning they do not rely on any specific programming language, and consist of HTML, CSS, and JavaScript code. This allows the system to learn and generate code snippets for a wide variety of web applications.

Data Evaluation - We meticulously gathered the code snippets, ensuring that only the most relevant ones were included. However, even with this careful selection, we identified some lines of code that were not utilized by the applications. To address this issue, we applied some code coverage strategies, rectified erroneous code, removed incomplete sections, eliminated unnecessary CSS, and replaced variable names that failed to describe their purpose. For instance, we replaced the variable “c” with “count” for improved clarity. Furthermore, we conducted a thorough manual evaluation of all the code samples.

This evaluation involved running the applications after the necessary modifications were made, allowing us to verify their correctness and usefulness. The main purpose behind this evaluation was to provide exemplary code examples that can serve as robust foundations for the described apps, with accurate code structures. We relied on the garbage in, garbage out (GIGO) concept that nonsense code will produce nonsense outputs.

Data Description - The corpus size is relatively small, consisting of 200 language descriptions and code pairs. We chose this size due to limitations in computing power and memory. However, there is potential to expand the dataset in the future by including additional samples in various languages. On average, the descriptions contain 15 words each. They are written in simple, non-technical language, avoiding excessive use of specialized terminology. The average number of lines in the HTML, CSS, and JavaScript files is 9, 17, and 9, respectively. Figure 1 presents an example of a natural language-code pair that will be further preprocessed.

Description			
Create an app called Notes App using HTML, CSS, and Javascript. The app will allow users to add, edit, and delete notes easily.			
HTML	CSS	JavaScript	
<pre> 1 <!DOCTYPE html> 2 <html lang="en"> 3 <head> 4 <meta charset="utf-8"> 5 <link rel="stylesheet" href="style.css" /> 6 <title>Notes App</title> 7 </head> 8 <body> 9 <button class="add" id="add">Add note</button> 10 <script src="script.js"></script> 11 <script src="https://cdn.jsdelivr.net/npm/marked/marked.min.js"></script> 12 </body> 13 </html> </pre>	<pre> 1 * { 2 box-sizing: border-box; 3 } 4 5 body { 6 background-color: #f0f0f0; 7 display: flex; 8 flex-wrap: wrap; 9 margin: 10px; 10 padding-top: 20px; 11 } 12 13 .note { 14 position: flex; 15 right: 10px; 16 background-color: #e0ffe0; 17 color: #008000; 18 border: none; 19 padding: 5px 10px; 20 margin-bottom: 10px; 21 cursor: pointer; 22 } 23 24 .note { 25 background-color: #fff; 26 margin: 10px 0 10px 10px; 27 height: 40px; 28 width: 400px; 29 overflow: scroll; 30 } 31 32 .note .tools { 33 background-color: #e0ffe0; 34 display: flex; 35 justify-content: flex-end; 36 padding: 2px 5px; 37 } 38 39 .note .tools button { 40 background-color: transparent; 41 border: none; 42 color: #008000; 43 cursor: pointer; 44 font-size: 16px; 45 margin-left: 5px; 46 } </pre>	<pre> 1 const addButton = document.getElementById("add"); 2 3 const addNewNote = (text = "") => { 4 const note = document.createElement("div"); 5 note.classList.add("note"); 6 note.innerHTML = ` 7 <div class="tools"> 8 <button class="edit">Edit / Save</button> 9 <button class="delete">Delete</button> 10 </div> 11 <div class="main">\${text ? "hidden" : ""}</div> 12 <textarea class="\${text ? "hidden" : ""}" /> 13 </div> 14 const editButton = note.querySelector(".edit"); 15 const deleteButton = note.querySelector(".delete"); 16 const main = note.querySelector(".main"); 17 const textarea = note.querySelector("textarea"); 18 textarea.value = text; 19 20 deleteButton.addEventListener("click", () => { 21 note.remove(); 22 // updateLocalStorage(); 23 }); 24 25 editButton.addEventListener("click", () => { 26 main.classList.toggle("hidden"); 27 textarea.classList.toggle("hidden"); 28 main.innerHTML = textarea.value; 29 }); 30 31 document.body.appendChild(note); 32 33 34 addButton.addEventListener("click", () => addNewNote()); </pre>	

Figure 1. Natural language – code pairs.

Data Preprocessing - We divided the dataset into three parts: 70% for training, 20% for validation, and 10% for testing. To facilitate data storage, we converted the data into a JSON file format. The code is stored as a single value, with specific separators used between the HTML, CSS, and JavaScript files: “\n-HTML-\n”, “\n-CSS-\n”, and “\n-JavaScript-\n”. Since the descriptions will come from spoken input, they do not include punctuation and are in lowercase. Therefore, we

made sure to lowercase the descriptions and remove all punctuation marks. To prepare our dataset for training, we utilized the Dataset class from HuggingFace datasets, which allows us to save a model with the following features: “description” and “code”.

3.2 CodeT5 Model Finetuning

In the introduction of our paper, we presented several popular architectures commonly used for language-related tasks. However, there are two main drawbacks associated with these existing methods. Firstly, many of these approaches rely on either an encoder-only model, such as BERT, or a decoder-only model, like GPT (which uses a transformer architecture with a stack of decoder layers). Unfortunately, these models are not specifically optimized to excel in both code generation and understanding tasks simultaneously. Secondly, most current methods treat source code as a sequence of tokens, similar to natural language, neglecting the valuable structural information present in programming languages. This structural information plays an important role in fully comprehending the semantics of the code (Epure and Iftene, 2016).

Architecture - Given the objective of our paper, we primarily focus on the T5 architecture, specifically using the CodeT5 (Wang *et al.*, 2022) model. This model, previously mentioned in (Wang *et al.*, 2021), is a pre-trained encoder-decoder model that has been specifically designed with an understanding of identifiers. CodeT5 achieves impressive performance across various code-related tasks, including understanding tasks like code defect detection and clone detection, as well as generation tasks involving programming language to natural language, natural language to programming language, and programming language to programming language. Considering our emphasis on the code generation task, we have selected the CodeT5 model due to its adaptability, impressive performance, and suitability for addressing the challenges inherent to code generation tasks.

Configuration - CodeT5 makes use of teacher forcing¹ during its training process, which entails providing both an input sequence and its corresponding target sequence. As CodeT5 is built upon the T5 architecture, it has the flexibility to be trained using both supervised and unsupervised approaches. However, for this paper, we will focus on the supervised training method. Supervised training involves utilizing labeled examples where the input sequences are paired with their corresponding target sequences. This approach allows the model to learn from the provided labeled data, aligning the input-output relationships effectively. We have used the T5ForConditionalGeneration model, which includes a language

¹ <https://towardsdatascience.com/what-is-teacher-forcing-3da6217fed1c>

head positioned above the decoder. In the supervised setup, our input sequence and output sequence follow the conventional sequence-to-sequence mapping commonly used in this framework. To enhance the model's performance for our specific task, we introduced a prefix to the input sequence: "generate application code:". This prefix matters because our task is similar to the code generation task used in CodeT5 pre-training. To achieve this, we set predefined maximum lengths for both the input and output sequences. Specifically, we defined the `max_input_length` and `max_target_length` as 128 and 1024, respectively. Any sequences exceeding these limits were truncated, while shorter sequences were padded to reach the designated lengths. By enforcing these maximum lengths and applying padding, when necessary, we create uniformity among the input and output sequences within each batch. This ensures compatibility during training and allows for efficient parallel processing. Additionally, truncating longer sequences helps to manage computational resources and mitigate potential memory limitations. Careful consideration was given when selecting these maximum lengths to strike a balance between capturing sufficient contextual information and maintaining computational efficiency. To ensure accurate loss calculation and prevent the padding tokens from influencing the training process, we took some important steps. In the labels, we replaced the padding token IDs with a special value of -100, known as the `ignore_index`. Thus, we instruct the loss function to ignore their contributions during computation. In addition, we supplied the `attention_mask` to the model. This `attention_mask` serves the purpose of highlighting which tokens should be attended to and which should be ignored. By incorporating the `attention_mask`, we effectively inform the model to ignore the padding tokens and focus solely on the relevant content. For optimizing the model's parameters during training, we employed the AdamW optimizer. To enhance the learning process, we set a slightly higher learning rate compared to the default value specified in the Trainer. Specifically, we utilized a learning rate of $3e-4$. The choice of a slightly elevated learning rate was made strategically to facilitate faster convergence and improved optimization of the model's parameters. By adjusting the learning rate, we aimed to strike a balance between the speed of learning and the stability of the training process. The AdamW optimizer further contributes to the optimization process by incorporating weight decay, enabling better generalization and reducing the risk of overfitting. During the inference process, we use the `generate()` function, which serves an important role in encoding the input and facilitating the generation of decoder outputs in an auto-regressive manner. The `generate()` function takes care of the entire encoding process, including the encoding of the input sequence and the propagation of the encoded hidden states through cross-attention layers to the decoder. This enables the model to generate output sequences based on the learned patterns and

contextual information. By using the *generate ()* function, we simplify the inference procedure, abstracting away the complexities of encoding and decoding. This allows for efficient and accurate generation of decoder outputs, ensuring the production of contextually relevant and semantically coherent code.

3.3 Integrating the Model into a Web Application

The code generation model was integrated into a web application developed using the Django framework. This web application includes two core components, namely the speech live transcription feature and the code generation functionality based on the transcribed description provided by the user. The speech live transcription component captures and transcribes the users spoken input in real-time. The transcribed description serves as input to the code generation component. Using our model, the application generates corresponding code snippets based on the transcribed description. This allows users to effortlessly transform their spoken ideas and requirements into functional code, simplifying the development process.

Speech to Text. The speech live transcription feature was implemented using the Deepgram API², WebSockets and channels³ within the Django framework. Deepgram’s API offers a powerful solution for transcribing audio and video files, catering to a wide range of file formats and languages. With its advanced capabilities, developers can easily integrate Deepgram’s API into their applications to extract accurate and reliable transcriptions. WebSockets establish a two-way connection between the client and the server, enabling real-time retrieval of speech input. This stateful protocol requires authentication only once, allowing for seamless data exchange between the client and server without delays. One of the advantages of WebSockets is their compatibility across different browsers, ensuring a consistent experience for users regardless of their preferred web browser. In the Django framework, the setup of WebSockets is a straightforward process. It is done through the integration of asynchronous Python and Django channels. To implement WebSockets in Django, we created an ASGI (Asynchronous Server Gateway Interface) server using Django channels. We then set a group where text messages could be sent in real-time, enabling instantaneous transmission of speech data. During the implementation process, we encountered version compatibility issues between Websockets and Django channels. To overcome this challenge, we opted for the channels==3.0.4 and django==4.0.0 versions, ensuring a stable and compatible environment for the WebSockets-based

² <https://deepgram.com/>

³ <https://channels.readthedocs.io/en/stable/>

speech live transcription feature. When the web application is launched, users are presented with a simple and user-friendly interface that features a prominent “Record” button. Upon clicking this button, users are prompted to grant permission for microphone access. This step ensures that the application can capture the user’s voice input accurately. Once the user has granted permission, they can begin describing the desired application using natural language. When the user finishes describing the application, they simply click the “Stop” button. This action signals the application to stop recording the voice input and proceed to the next step. After clicking the “Stop” button, the user interface automatically reveals the “Generate Code” button. This button serves as a clear call to action, prompting the user to initiate the code generation process based on the transcribed description.

Text to Code. The code generation component of the web application is built upon our fine-tuned model, which has been integrated into the system using the Joblib library⁴. Using the Joblib for storing and loading machine learning models offers several advantages over alternative techniques. One important benefit of using Joblib is its efficient storage mechanism. Data is stored as byte strings rather than objects, enabling quick and space-efficient storage. Compared to traditional pickling methods, Joblib can save models in a smaller amount of space, allowing for more efficient resource utilization. Additionally, Joblib provides built-in error handling when reading or writing files. This automatic error correction mechanism improves the robustness of model storage and retrieval, reducing the risk of data corruption. Moreover, Joblib supports multiprocessing across multiple machines or cores on a single machine. This enables programmers to parallelize tasks and distribute them across numerous machines, clusters, or GPUs. Once the user clicks the “Generate Code” button, the web application initiates the code generation process. It is important to note that generating the code may take a few moments to complete, depending on the complexity of the application being described. During this short waiting period, users will be provided with a visual indication - a loading text animation, to signify that the code is being generated. This feedback ensures that users are aware that the system is actively processing their requests. After a few moments, the generated code will be displayed on the screen, conveniently separated into HTML, CSS, and JavaScript sections (Figure 2). Each section will be clearly labeled, allowing users to easily identify and understand the purpose of each code segment. The displayed code is fully editable, providing users with the flexibility to make any necessary modifications according to their specific requirements. Users can copy the code directly from

⁴ <https://joblib.readthedocs.io/en/stable/>

the screen and paste it into their preferred code editor or development environment for further refinement or integration into their projects.



Figure 2. Web Application Interface – The Generated Code.

4 Evaluation

This section provides a comprehensive evaluation of the performance and effectiveness of our code generation model developed in this study. In this section, we present a detailed analysis of the model’s performance metrics, including BLEU score, ROUGE score, and F1 score, previously mentioned in Section 3.2. Additionally, we provide a qualitative evaluation of the generated code outputs, examining their alignment with the natural language descriptions of the application. The evaluation was conducted on a carefully curated dataset described in Section 3.1. Furthermore, we present some case studies that showcase the model’s performance in real-world scenarios. These case studies demonstrate the model’s capability to accurately capture the intended functionality. Additionally, we conduct an error analysis, identifying common challenges and limitations encountered during the code generation process. Quantitative results highlight the model’s proficiency in generating code across HTML, JavaScript, and CSS languages. We report the achieved BLEU score of 0.26, indicating a measure of code similarity between the generated code and human-written references. Additionally, the model obtained a ROUGE score of 0.28, illustrating the quality of the generated code in capturing important features mentioned in the natural language descriptions. Furthermore, the F1 score, computed using the formula mentioned in Section 3.2, yielded a value of 0.27, indicating the overall effectiveness of the code generation model. It is worth highlighting that these

results are even more remarkable when considering the limited size of the dataset. Despite the challenges posed by a small dataset, our model showcases its capability to generate code effectively. We also conducted a comparative analysis between our generated code and the responses obtained from ChatGPT free chat, aiming to assess the usability and practicality of our code generation model. This evaluation sought to provide insights into the quality and effectiveness of the generated code in comparison to an alternative approach. To perform this comparison, we selected a set of case studies representing various scenarios and application requirements. For each case study, we provided the same natural language description to both our code generation tool and the ChatGPT free chat system. We then analyzed and compared the code outputs generated by both approaches. The comparison focused on several key aspects, including code accuracy and overall usability. We carefully examined the generated code from both sources, considering factors such as code readability, efficiency, and suitability for the intended application.

4.1 Case Studies

To test the performance and effectiveness of our code generator tool, we conducted a user test where we asked a participant to provide a brief description of their desired web application. The participant spoke the description as follows: *“Create a web application that will allow users to flip a coin on a button click, and the web app will display the result”*. Our tool generated code snippets in HTML, CSS, and JavaScript based on the provided description. While the generated HTML and CSS code successfully captured the requirements and maintained syntactic correctness, it encountered challenges in generating syntactically correct JavaScript code. The results can be seen in Listings 1., 2., and 3.

Listing 1: HTML Generated Code.

```
<!DOCTYPE html>
<html>
  <head> <meta charset = "utf-8">
    <title>Coin Flip</title> <link rel = "stylesheet" href =
      "style.css"/>
  </head>
  <body> <div class = "container">
    <div id = "result"></div> <button class = "btn" id =
      "flipCoinBtn">Flip Coin</button>
    <script src = "script.js"></script>
  </body></html>
```

Listing 2: CSS Generated Code.

```
body {
  background-color: whitesmoke; text-align: center; flex-direction:
column;
  display: flex; justify-content: center; align-items: center; height:
100vh;}
.container{
  padding: 10px;}
.btn{
  padding: 10px; background-color: green;
  color: white; border: none; border-radius: 6px;}
```

Listing 3: JavaScript Generated Code.

```
const flip = document.getElementById('flipCoinBtn');

const flipCoinBtn = document.getElementById('flipCoinBtn');
const flipCoinBtn = document.getElementById('flipCoinBtn');
const flipCoinBtn.addEventListener('click', ()=> {let coin =
coin.flip();
  let coin = coin.flip(); let coin = coin.flip(); let coin =
coin.flip();
  let coin = coin.flipCoin(); let coin = coin.flipCoin();})
```

Despite the limitations in the JavaScript code generation, we observed that the generated code had a clear structure and organization, making it easily understandable and modifiable. In terms of functionality, the generated code successfully implemented the desired coin-flipping feature and accurately displayed the result upon button click, as outlined in the user's description. This demonstrates the capability of our code generator tool to capture and translate user intentions into working code. Although improvements are needed in the generation of syntactically correct JavaScript code, the overall output showcases the potential of our code generation tool to assist developers in quickly generating the entire web application code. For the second case study, we conducted an experiment asking our tool to generate code for an application that was not included in our training dataset. Specifically, we asked the tool to: “*Generate a body mass index calculator application that calculates body mass index based on height and weight inputs*”. As anticipated, since our model was not explicitly trained on this specific application, the generated code did not demonstrate full functionality. However, it did demonstrate an understanding of the syntax by correctly generating the title and input fields required for height and weight. This outcome highlights the limitations of our current model and dataset in capturing the complexity of every possible application scenario. It also emphasizes the importance of training on a diverse range of examples to enhance the tool's ability to generate functional code for a wider range of applications.

4.2 Automated vs Human Evaluation

While automated evaluation metrics like BLEU and ROUGE-N scores provide valuable insights into the quality of the generated code, they may not fully capture its usability aspects. To bridge this gap and gain a deeper understanding of the code's usability, we conducted a human evaluation in a code environment. To evaluate the usability of the generated code, we set up a controlled testing environment where the generated code was executed and examined for its functionality. The code was tested in a code environment, allowing us to interact with the web application and check its behavior. We carefully observed how the application responded to user inputs, such as button clicks, and verified that the expected functionalities, as described in the user's initial input, were successfully implemented. By performing this careful evaluation, we aimed to validate the usability and functionality of the generated code. The results of our human evaluation demonstrated that with minimal modifications to the code, the generated web application became fully functional. The inclusion of human evaluation in our strategy adds an important layer of insight that complements the quantitative metrics provided by automated evaluation methods. It allows us to gain an extensive understanding of the code's performance in real-world scenarios and evaluate its practical applicability. The initial results obtained from this short evaluation provide valuable insights into the usability of the generated code and highlight its potential as a time-saving tool for web application development.

Table 1. ChatGPT vs Our Tool.

ChatGPT	Our Tool
Trained on a large dataset	Trained on limited data
Sometimes only provides explanations for tasks and does not generate any code	Always generates some code
Wider range of domain knowledge	Limited domain knowledge
May not capture the entire underlying context	May not capture the entire underlying context
May generate the code in a single file	Always will separate the logic in separate files
Does not use semantic tags (many times it uses the <div> element by default)	Also uses semantic tags such as <fieldset>
Does not adapt so easily to new syntax patterns	Adapts easily to new syntax patterns
Needs big amounts of data for fine-tuning	Achieves good performance with relatively smaller amounts of fine-tuning data
May not generate functional code	May not generate functional code

4.3 Viable ChatGPT Alternative for Writing Code

We also conducted a comparison analysis by providing the same application description to ChatGPT, a language model trained on a large dataset. It is important to note that ChatGPT has been exposed to a vast amount of training data, which makes it a powerful language model. As expected, the code generated by ChatGPT and our code generation tool differed due to the different architectures and training data. However, both approaches successfully captured the functionality requested by the user in the first case study and produced functional HTML and CSS code. Notably, our code generation tool excelled in generating more sophisticated CSS, adding additional style elements to the application. While ChatGPT managed to generate the correct JavaScript code in this specific case, it is important to consider the strengths and limitations of each approach. Our tool, despite being trained on a smaller dataset, was able to deliver noteworthy results, capturing the essential functionality and generating code that fulfilled the user's requirements. Another notable aspect of our model is its ability to capture syntactic information, despite being trained in separate languages and having no specific training in HTML and CSS. This demonstrates the effectiveness of the CodeT5 architecture in handling diverse coding languages and adapting to new syntax patterns. One significant advantage of the CodeT5 architecture we observed is its ability to achieve good performance with relatively smaller amounts of fine-tuning data. This characteristic not only saves computational resources but also reduces the overall training time. Gathering sufficient data for training can be a time-consuming task, and the ability to achieve satisfactory results with less data is a valuable advantage. It is important to acknowledge that the comparison between our tool and ChatGPT serves to highlight the different characteristics of each approach rather than determine superiority. Both tools have their strengths and weaknesses, summarized in Table 1, and the generated code should be evaluated based on its usability, readability, and compliance with the desired functionality. Given the limitations of our tool's training dataset, we consider the achieved results to be satisfactory. The ability to generate functional code, particularly in HTML and CSS, is a promising outcome.

5 Usability Testing

To further evaluate the usability of the speech-based code generation tool, a series of evaluations were done by colleagues from "Alexandru Ioan Cuza" University of Iasi. The participants were selected based on their familiarity with programming languages and their expertise in software development.

Evaluation Procedure - Each participant was given a brief introduction to the speech-based code generation tool, along with a set of predefined tasks that

represented common coding scenarios. They were instructed to use the tool to generate code for applications of their liking.

Feedback and Observations - Upon completion of the evaluation sessions, participants were asked to provide their feedback on various aspects of the tool. Their responses were analyzed to identify common problems and patterns.

Effectiveness - The participants generally found the speech-based code generation tool to be effective in generating applications accurately for simpler applications. They reported that the tool accurately translated their spoken description into functional code. This indicates that the code generation techniques employed in the tool were successful in achieving the desired functionality.

Usability - In terms of usability, the tool received positive feedback from the participants. They appreciated the simplicity of the user interface and found the interaction with the tool to be intuitive. It is worth noting that while most participants indicated satisfaction with the accuracy of the speech transcription, a few mentioned that their evaluation might be affected by English not being their first language. They acknowledged the possibility of mispronouncing certain words and expressed a slight uncertainty in evaluating the transcription accuracy in such cases.

Limitations and Suggestions for Improvement - Despite the overall positive feedback, a few limitations and areas for improvement were identified by the participants. Some reported instances where the tool struggled to accurately interpret more complex or ambiguous descriptions, leading to unexpected code outputs. Additionally, participants suggested highlighting the code snippets for each language for better understanding and visualization. In conclusion, the evaluation by our colleagues demonstrated that the speech-based code generator tool is a promising tool in facilitating the coding process. The participants were enthusiastic about the tool and reported that they would use such a tool in their daily tasks. However, there are still some areas that require further refinement, such as handling complex descriptions and complex application code. These findings will serve as valuable insights for future iterations of the tool.

6 Conclusions

The primary objective of this paper was to explore the possibility and feasibility of designing a speech-based code generation system capable of generating not just code in a specific language, but the entire application code, particularly in HTML, CSS, and JavaScript. To improve the development process, an additional component was incorporated to transcribe the spoken descriptions. These days, programmers face limitations in terms of computational resources for complex

tasks like code generation. Consequently, adapting a pre-trained model to a new task became crucial, allowing them to work within resource constraints, reduce training time, and enhance model generalization. Among the most time-consuming and challenging aspects of our research were identifying the optimal model for the task and performing data collection, along with its manual evaluation. Careful consideration was given to configuring the model and selecting appropriate hyperparameters, which significantly contributed to achieving improved results.

In the future, one important area to focus on is expanding the dataset to include a wider range of examples, allowing the model to learn from a more diverse set of code patterns. This would improve our tool's capabilities and potentially apply it in various domains beyond web development, such as mobile application development or other programming projects.

References

- Allamanis, M., Barr, E. T., Devanbu, P., and Sutton, C. 2018. A Survey of Machine Learning for Big Code and Naturalness. *ACM Comput. Surv.*, 51(4).
- Aronowitz, H., Gat, I., Morais, E., Zhu, W., and Hoory, R. 2022. *Towards a Common Speech Analysis Engine*. <https://arxiv.org/abs/2203.00613>
- Begel, A. 2005. Programming by Voice: A Domain-Specific Application of Speech Recognition. *BSc. Thesis, Programvarutechnik*. <https://api.semanticscholar.org/CorpusID:16889916>
- Epure, C. and Iftene, A. 2016. Semantic Analysis of Source Code in Object-Oriented Programming. A Case Study for C#. *Romanian Journal of Human - Computer Interaction*, 9(2):103–118.
- Gu, X., Zhang, H., Zhang, D., and Kim, S. 2016. Deep API Learning. *CoRR*, *abs/1605.08535*.
- Gvero, T. and Kuncak, V. 2015. Synthesizing JAVA Expressions from Free-Form Queries. *SIGPLAN Not.*, 50(10):416-432.
- Husain, H., Wu, H.-H., Gazit, T., Allamanis, M., and Brockschmidt, M. 2019. Codesearchnet Challenge: Evaluating the State of Semantic Code Search. *CoRR*, *abs/1909.09436*.
- Khan, F., Chiarcos, C., Declerck, T., Gifu, D., González-Blanco García, E., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J., Pagé-Perron, É., Passarotti, M., Ros, S., Truică, C.-O. 2022. When Linguistics Meets Web Technologies. Recent Advances in Modelling Linguistic Linked Open Data. *Semantic Web Journal*, Tracking #: 2859-4073, 13(6):987-1050.
- Kushman, N. and Barzilay, R. 2013. Using Semantic Unification to Generate Regular Expressions from Natural Language. In *Proceedings of the 2013 Conference of the North American Chapter of the ACL: HLT*, 826-836.
- Lin, X. V., Wang, C., Pang, D., Vu, K., Zettlemoyer, L., and Ernst, M. D. 2017. Program Synthesis from Natural Language Using Recurrent Neural Networks. *Technical Report*

- UW-CSE-17-03-01, University of Washington Department of Computer Science and Engineering, Seattle, WA, USA, March.*
- Ling, W., Grefenstette, E., Hermann, K. M., Kocisky, T., Senior, A. W., Wang, F., and Blunsom, P. 2016. Latent Predictor Networks for Code Generation. *CoRR*, *abs/1603.06744*.
- Maloku, R. S. and Pllana, B. X. 2016. Hypercode: Voice Aided Programming. *IFAC-PapersOnLine*, 49(29):263–268. 17th IFAC Conference on International Stability, Technology and Culture TECIS 2016.
- Oda, Y, Fudaba, H., Neubig, G., Hata, H., Sakti, S., Toda, T., and Nakamura, S. 2015. Learning to Generate Pseudo-code from Source Code Using Statistical Machine Translation. In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 574-584.
- Quirk, C., Mooney, R., and Galley, M. 2015. Language to Code: Learning Semantic Parsers for if-this-then-that Recipes. In *Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP 1*, 878-888.
- Rosenblatt, Lucas. 2017. Vocalide: An Idea for Programming via Speech Recognition. In *Proceedings of the 19h International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '17, page 417-418.
- Sun, Z., Zhu, Q., Xiong, Y., Sun, Y., Mou, L., and Zhang, L. 2020. Treegen: A Tree-Based Transformer Architecture for Code Generation. *Proceedings of the AAAI*, 34(05):8984-8991.
- Wang, Y., Wang, W., Joty, S., and Hoi, S. C. H. 2021. *Codet5: Identifier-Aware Unified Pre-trained Encoder-Decoder Models for Code Understanding and Generation*. <https://arxiv.org/abs/2109.00859>
- Wang, Y., Gotmare, A., Savarese, S., and Hoi, S. 2022. Coderl: Mastering Code Generation through Pretrained Models and Deep Reinforcement Learning. *CoRR arXiv:2207.01780*, 07.
- Yin, P. and Neubig, G. 2017. A Syntactic Neural Model for General-Purpose Code Generation. *CoRR*, *abs/1704.01696*.
- Zhong, V., Xiong, C., and Socher, R. 2017. Seq2sql: Generating Structured Queries from Natural Language Using Reinforcement Learning. *CoRR*, *abs/1709.00103*.
- Trandabăţ, D. and Gifu, D. 2017. Social Media and the Web of Linked Data. *Proceedings of 2017 ACM/IEEE Joint Conference on Digital Libraries*, JCDL, Toronto, pages 350-351, IEEE Catalog Number: CFP17JCD-ART (Xplore). DOI: 10.13140/RG.2.2.30786.89285.
- Țucă, L.-P. and Iftene, A. 2017. Speech Recognition in Education, Voice Geometry Painter Application. *9th Conference on Speech Technology and Human-Computer Dialogue*, IEEE, pages 1-8.

INVESTIGATION OF ROMANIAN SPEECH RECOGNITION IMPROVEMENT BY INCORPORATING ITALIAN SPEECH DATA

CAROL-LUCA GASAN¹, VASILE PĂIȘ²

*¹Research Institute for Artificial Intelligence “Mihai Drăgănescu”,
Romanian Academy
gasancarolluca@gmail.com*

*²Research Institute for Artificial Intelligence “Mihai Drăgănescu”,
Romanian Academy
vasile@racai.ro*

Abstract

This study investigates the potential for augmenting Romanian automatic speech recognition (ASR) systems by incorporating Italian speech data, leveraging the phonetic affinity shared between the two languages. Methodologically, we curate a unique dataset comprising Italian audio aligned with Romanian text, facilitated by grapheme-to-phoneme (G2P) and phoneme-to-grapheme (P2G) models. Furthermore, this work represents a foundational underpinning for forthcoming investigations involving combinations of other phonetically related languages. It offers a perspective into the prospects of harnessing phonetic commonalities to improve ASR technology especially in underrepresented categories of speech and hardware-restricted conditions.

1 Introduction

Automatic Speech Recognition (ASR) technology plays a pivotal role in our increasingly interconnected world, enabling machines to decipher spoken language and convert it into textual form. As the demand for accurate ASR systems continues to grow, so does the need to enhance their performance across diverse linguistic landscapes and hardware specifications.

This paper focuses on proposing a method that undertakes the lack of resources for the Romanian language compared to other more developed languages and as well the cases where an underpowered hardware configuration is of necessity, such as mobile applications and voice assistants. The main objective of this article is to integrate resources from a phonetically similar language, in this case Italian,

into the development of a Romanian ASR model and analyse how it affects the overall performance. This choice is motivated by the more resourceful character of Italian (according to the European Language Grid¹ it has around 60% more language datasets than Romanian, as seen in Figure 1), and by the similarity of the two languages (Figure 2), which makes the included data play the role of generalising the voices and accents within the overall composed dataset. Yet it should be noted that the similarity of two languages is subject to dialectal variations, which might influence the model’s performance.

However, it is imperative to acknowledge that this is as well a work-in-progress, representing an important step towards a broader understanding of the complex interplay between Romance languages in ASR. As we present in Section 4, we intend to continue by integrating other languages as well (such as Spanish, French and Portuguese), and to perfect the data conversion pipeline to increase its accuracy and quality. The rest of this paper is structured as follows: Section 2 presents related work, Section 3 introduces our method, and we conclude in Section 4.

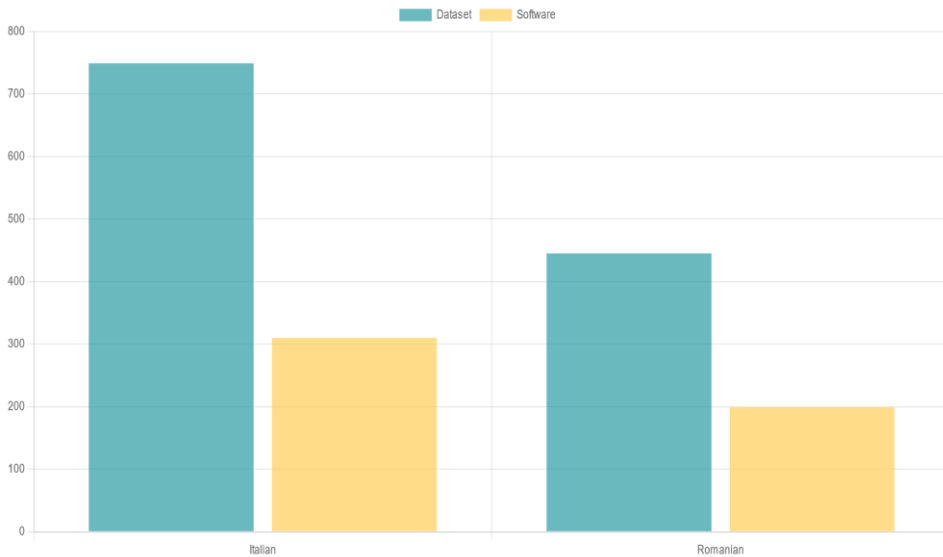


Figure 1. Romanian and Italian resources’ count according to the European Language Grid.

¹ <https://live.european-language-grid.eu>

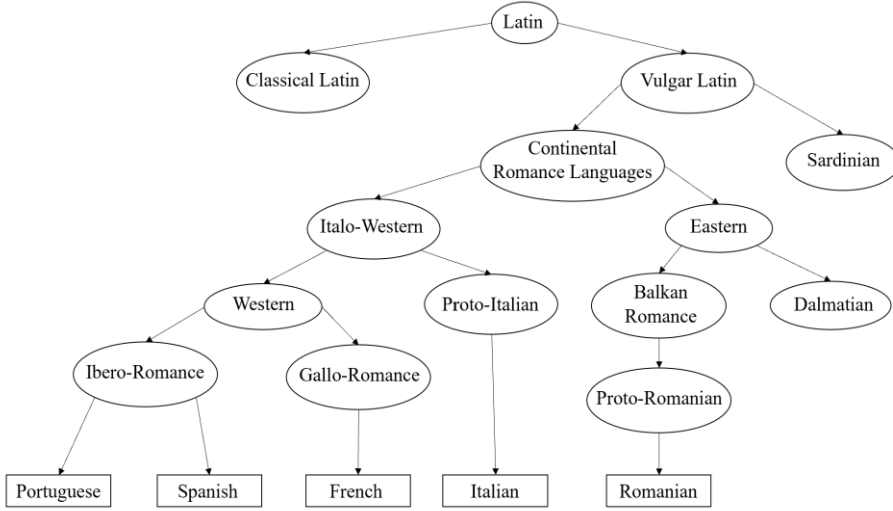


Figure 2. A possible romance languages classification.

2 Related Work

2.1 Romanian ASR and Resources

Speech corpora for the Romanian language includes among other resources the following datasets: ROBIN Technical Acquisition Speech Corpus – RTASC (Păiș *et al.*, 2021), Romanian Read-Speech Corpus – RSC (Georgescu *et al.*, 2020), Romanian Speech Synthesis Corpus – RSS (Stan *et al.*, 2011), MARA (Stan *et al.*, 2021), SWARA (Stan *et al.*, 2017). The Representative Corpus of Contemporary Romanian Language – CoRoLa (Barbu-Mititelu *et al.*, 2014), even though it presents most of its data in textual form, it also offers speech data adding to around 300 hours. The Bimodal Corpus for Romanian Language – CoBiLiRo (Cristea *et al.*, 2020) is a research project part of ReTeRom², which provides different annotation systems along with textual and audio data.

USPDATRO (Păiș *et al.*, 2023a; Păiș *et al.*, 2023b) is a project meant to gather underrepresented Romanian speech. Therefore, it contributes to the diversity characteristic of the dataset, making it a very good resource for a fair evaluation of ASR models. The authors of the aforementioned paper trained a Whisper model that achieves a WER of around 10% on average in a certain environment and showed that when tested on the USPDATRO it only obtains a 28% WER, proving the lack of underrepresented speech for the Romanian language.

² <https://www.racai.ro/p/reterom/>

The Common Voice initiative (Ardila *et al.*, 2020) is based on crowdsourcing and offers speech data for approximately 40 languages, including Romanian. It's a resource that is updated every three months and new languages are still being introduced. Vox Populi (Wang *et al.*, 2021) is a multilingual corpus containing 100K hours of unlabelled speech in 23 languages. Another multilingual dataset is FLEURS (Conneau *et al.*, 2022) with around 12 hours of speech supervision per language, including Romanian. Part of the LibriVox³ project, a volunteer effort to provide free of licence audiobooks, is the LibriSpeech dataset (Panayotov *et al.*, 2015) providing a total of 1k hours of audiobooks, as well in Romanian.

Georgescu *et al.* (2021) include in their work datasets such as SSC1, SSC2 and CDP that were used during the experiments only as the evaluation splits offered, and they provided an achievement of 1.6% WER on the RSC dataset. Avram *et al.* (2021) and (2020) present trainings using the DeepSpeech2 (Amodei *et al.*, 2016) model and obtain a WER of 9.9% on a mixed evaluation set. The authors of (Avram *et al.*, 2023) experiment with the Wav2Vec2 architecture (Baevski *et al.*, 2020) and report a WER of 1.78% on RSC.

Table 1. Chosen Romanian speech datasets for training.

Dataset	Speech Type	Domain	# Hours	# Utterances	# Speakers
CoRoLa	Read, Spontaneous	News, Radio	152	-	-
RSC	Read	Wikipedia	100	136.1k	164
VoxPopuli	Spontaneous	Legal	83	27k	164
CV	Read	Wikipedia	9	8k	130
RTASC	Read	Technology	6.5	3.8k	6
RSS	Spontaneous	Internet, TV	5.5	5.7k	3
USPDATRO	Read, Spontaneous	Internet	4.25	2.6k	84

2.2 Same Language Family ASR

Zgank (2019) presents an experiment regarding the development of speech recognition models for languages that are part of the same family group, in this case Spanish and Romanian. It also presents the mapping between the two languages' phonemes and the difficulties encountered in providing the mapping, while the final model is at word-level (every utterance consists of a single read word) and has the purpose of comparing two proposed mapping methods.

Cerniavski and Stymne (2023) further support the benefits of sharing resources within the same language family, where it treats in detail all the possible combinations of data between Swedish, Danish, and Norwegian in order to train a speech recognition model for Swedish.

³ <https://librivox.org/>

2.3 Phonetic Dictionaries

Wiktionary⁴ is a multilingual dataset of phonetic transcriptions which are manually crafted as a collaborative initiative motivated by the lack of resources. The phonemes respect the rules of the International Phonetic Alphabet – IPA⁵. The Italian CV dictionary (Ahn and Chodroff, 2022) was released as an initiative to obtain better alignment accuracy and it contains around 67k entries in IPA.

MaRePhor (Toma *et al.*, 2017) stands for Machine-Readable Phonetic Dictionary for Romanian and provides around 73k words and their phonetic form using the Speech Assessment Methods Phonetic Alphabet – SAMPA⁶. NaviRO⁷ is a dataset introduced in the paper (Domokos *et al.*, 2014) which takes on the problem of creating a Romanian grapheme-to-phoneme model. This resource contains around 140k entries in SAMPA format. RoLEX (Lőrincz *et al.*, 2023) is an exclusively Romanian dictionary which offers the phonetic representations of around 330k words using the Extended Speech Assessment Methods Phonetic Alphabet – X-SAMPA⁸ format.

Table 2. Comparison of phonetic dictionaries for Romanian and Italian.

Dataset	# Ro Entries	# It Entries	Phonetic Alphabet
Wiktionary	63k	30k	IPA
It CV dict	-	67k	IPA
MaRePhoR	73k	-	SAMPA
NaviRO	140k	-	SAMPA
RoLEX	330k	-	X-SAMPA

2.4 G2P and P2G models

RECOApy (Stan, 2020) is a tool meant to help developing speech-based applications, through which they undertake the problem of training G2P models for multiple languages including Italian and Romanian. They mainly opted for the Wiktionary dataset, and for English and Romanian they also used the CMUDict⁹ and MeRaPhoR datasets respectively.

In the solving of the SIGMORPHON 2020 Task 1 by the University of Alberta (Hauer *et al.*, 2020), a broader perspective on the training of P2G and G2P models is chosen, by showcasing the differences between high data resources and low

⁴ <http://www.wiktionary.org/>

⁵ <https://www.internationalphoneticassociation.org/content/handbook-ipa>

⁶ <https://www.phon.ucl.ac.uk/home/sampa/>

⁷ <http://users.utcluj.ro/~jdomokos/naviro/>

⁸ <https://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>

⁹ <https://github.com/cmuspinx/cmudict>

data resources trainings of these models. The full dataset used for the high-resources context consists of around 40k samples for each language, which were extracted from the Wiktionary dataset using the command-line tool WikiPron¹⁰ (Lee *et al.*, 2020).

2.5 Linguistics Resources

The Romance Phonetics Database – RPD¹¹ contains documentation on 5 languages including Italian and Romanian and it also provides audio files of an example text read by multiple readers. The project is meant to clarify different confusions at phonetic level. SRoL¹² is a corpus representing an online dictionary of sounds and words for the Romanian language while also presenting linguistics particularities.

3 Method

Since the purpose of this paper is the development of a monolingual model and not a bilingual one, the Italian data should be transformed into a type of data that shares important characteristics with the Romanian data. Obviously, this operation should be split into two important parts: the conversion of the audio files and the conversion of the text files. At audio level, since a variety of voices is meant to improve the performance of the overall model, the only characteristic that might be relevant to change is the speed of speech. Table 3 shows statistics on the Common Voice dataset regarding this and it can be drawn that the data is sufficiently similar (at least on the train splits) such that no further action is needed on this side.

Table 3. Speed of speech in words-per-minute for the Common Voice dataset.

Language	Train Split	Test Split	Dev Split
Romanian	112.0	105.5	114.3
Italian	113.2	93.7	97.0

However, for the text files, there are two major components to be changed, one clearly visible and one harder to interpret. The first one is related to the alphabet and the characters specific to the languages. For example, the Italian character 'ó' doesn't exist in Romanian, and the Romanian diacritics ("ăĂ", "âÂ", "îÎ", "șȘ", "țȚ") are a sign of correctly written text. Moreover, diacritics play a vital role in the semantics of the words ("față" means *face*, while "fata" means *the girl*), in

¹⁰ <https://pypi.org/project/wikipron/>

¹¹ <https://rpd.artsci.utoronto.ca/>

¹² http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/en/arhiva_tot_en.htm

the grammatical structure of the sentences (“*fată*” refers to a generic girl, while “*fata*” refers to a specific girl, marking the presence of a definite article) and in the pronunciation of the given text (“ă” has associated the phoneme “ə”, while “a” has associated the phoneme “a”). On the other hand, the second component is related to the informally described “sound of the text”, which refers to the way a native speaker would read a certain sequence of characters while ignoring its lexical correctness. This introduces into the discussion the concepts of grapheme and phoneme, representing the way a text is written (the visual aspects) and respectively the way a text is read or pronounced (the phonetic aspects). Therefore, our task reduced to keeping the phonemes as similar to the initial ones (because they are a representation of the information kept in the spectrograms) and based on those resulted phonemes, to determine an accurate graphic representation. It should be mentioned that besides the Italian to Romanian conversion, there was no additional pre-processing of the text or audio data.

3.1 Conversion of the Italian Data

To achieve the aforementioned, there would be needed a G2P model for Italian, a P2G model for Romanian, and a mapping between the two sets of phonemes (as seen in Figure 3).

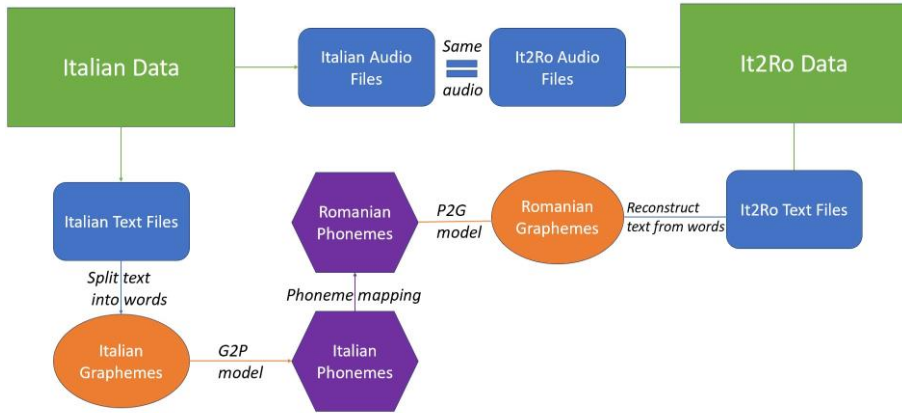


Figure 3. Processing Pipeline of the Italian Data.

For the G2P model, the dataset chosen for analysis is the Italian CV dictionary previously mentioned, which has 66,879 entries. For the P2G model, the chosen dataset is RoLEX with 331,250 entries. The reason behind this choice consists of their extensive size compared to other resources, which can prove to be useful in terms of so-called knowledge, keeping in mind that the models are going to be fed a very big diverse dataset. The two phoneme sets associated with the datasets in discussion and the mapping between them are captured in Table 4.

Table 4. Romanian and Italian Phonemes.

Phoneme Set for RoLEX (IPA)																		
.	a	b	c	d	e	ẽ	f	g	h	i	j	je	k	ks	k'	l	m	n
o	õ	p	r	s	t	ts	tʃ	u	v	w	z	ə	g	gz	g3	ɨ	ʒ	3

Phoneme Set for Italian CV Dictionary (IPA)																			
a	a:	b	b:	d	d:	d3	d3:	e	e:	f	f:	g	h	i	j	k	k:	l	l:
m	m:	n	n:	o	o:	p	p:	r	r:	s	s:	t	t:	ts	tʃ	tʃ:	u	v	v:
w	x	y	z	ä	æ	ð	ø	h	œ	ə	ə	ɛ	g	ɲ	ʃ	ʌ	ẽ		

Phoneme Mapping (IPA)																				
IT	a	a:	b	b:	d	d:	d3	d3:	e	e:	f	f:	g	h	i	j	k	k:	l	l:
RO	a	a	b	b	d	d	j	j	e	e	f	f	g	h	i	j	k	k	l	l

IT	m	m:	n	n:	o	o:	p	p:	r	r:	s	s:	t	t:	ts	tʃ	tʃ:	u	v	v:
RO	m	m	n	n	o	o	p	p	r	r	s	s	t	t	ts	tʃ	tʃ	u	v	v

IT	w	x	y	z	ä	æ	ð	ø	h	œ	ə	ə	ɛ	g	ɲ	ʃ	ʌ	ẽ		
RO	v	ks	i	z	i	ẽ	z	e	h	e	o	ə	e	g	j	f	l	i		

It should be noted that all the phonemes were used in the IPA format, and a conversion for RoLEX was needed from X-SAMPA to IPA (which is an exact conversion since every X-SAMPA phoneme has a single correspondent in IPA). Also, on top of the phoneme mapping, another small set of transformations was applied to increase the accuracy of the phoneme conversion and eliminate common mistakes such as the ones in Table 5.

Table 5. Supplementary Transformations

Old Sequence	New Sequence	Note
gh	g ⁱ	-
ie	je	-
ch	k ⁱ	-
oa	õa	-
i	ĩ	The character should be at the end of the word

3.2 G2P and P2G Models

For training the models, we considered the CMUSphinx's sequence-to-sequence G2P¹³ architecture and modified it in order to obtain a new model specific to this work. The tool uses a transformer model from tensor2tensor toolkit rather than the usual recurrent neural networks approach. Also, for the Romanian model, the dataset was inverted such that the input is the word formed by the phonemes while the output is the word formed by the graphemes, therefore obtaining a P2G model. Some example entries can be seen in Table 6.

¹³ <https://github.com/cmusphinx/g2p-seq2seq#sequence-to-sequence-g2p-toolkit>

Table 6. Example Entries per Dataset State

Dataset	Example Entries	
RoLEX	ălălalt	@l@lalt
	şablonarzi	Sablonarzi_0
	puteaţi	pute_Xatsi_0
RoLEX-IPA	ălălalt	ələlalt
	şablonarzi	ʃablonarzi
	puteaţi	put̪eats̪i
RoLEX-IPA-inverted	ələlalt	ălălalt
	ʃablonarzi	şablonarzi
	put̪eats̪i	puteaţi
Italian-CV-dictionary	discussa	difus: a
	lignaggio	lipa dʒ: o
	schematizzato	skematitsato

The models were trained using an undefined number of epochs which corresponds to a training that stops when no further improvement in performance is noted. Though, to simulate a hardware-limited ecosystem, the training was stopped after 6 hours. The achieved performance was noted in Table 7 (where WER = word error rate, CER = character error rate and PER = phoneme error rate).

Table 7. Metrics for the trained G2P and P2G models.

Model Type	Dataset	WER	CER/PER
G2P	Italian-CV-dictionary	4.33%	0.80%
P2G	RoLEX-IPA-inverted	3.32%	0.67%

It should be noted that the system cannot be properly evaluated on another dataset such as Wiktionary and then compared, since the set of phonemes might be different for each dataset and language (even though the alphabet coincides sometimes). But for a quality check, short trainings were performed within the same environment on the Wiktionary dataset and it was drawn that they provide similar results (Table 8) to those obtained during the experiments presented in (Stan, 2020).

Table 8. Training experiment proves the efficiency of the model's architecture.

Model Type	Architecture	Dataset	WER	PER
G2P (current paper)	Transformer	Wiktionary (Italian)	18.75%	4.31%
G2P (Stan, 2020)	Transformer	Wiktionary (Italian)	18.67%	4.44%

3.3 ASR Model Choice

For the speech recognition model, we decided to finetune Whisper (Radford *et al.*, 2023). The model is a transformer-based encoder-decoder model renowned for its strong ability to generalise to different domains. We were interested in a model that can be employed at runtime in a hardware restricted environment, such as a phone or tablet, without the need for a powerful GPU. Thus, we decided to continue with the small checkpoint of the model that has 224M parameters. Furthermore, choosing a model with a reduced number of parameters helps in reducing the time required for training and testing the model. This contributes to reducing the carbon footprint associated with the processes, in accordance with the Green AI principles (Strubell *et al.*, 2019).

3.4 Initial Experiment

An initial experiment settled only on training with the Common Voice data to get a sense of the quantity of It2Ro data to be added. CV for Romanian has 5k entries for the train split while CV It2Ro has 15k entries for the dev split. Also, no It2Ro data was added to the overall validation split since this fact might be reflected during the evaluation which is done on lexical correct texts. Therefore, from the first experiments, Table 9 had been formed where ALL1 represents the eval sets of RTASC, RSC, CDP, SSC1 and SSC2, while ALL2 represents the eval sets of RSC, RSS and CoRoLa.

Table 9. Best results during the CV-only experiment.

It2Ro	LR	Epoch	CV Ro		ALL1		ALL2		USPDATARO	
			WER%	CER%	WER%	CER%	WER%	CER%	WER%	CER%
-	5e-6	10	25.99	7.92	43.15	18.34	31.56	12.48	46.33	21.28
1k	5e-6	8	27.93	8.40	40.96	16.25	30.74	11.27	46.24	19.36
1.5k	5e-4	3	26.72	7.74	45.29	19.45	40.33	20.08	51.62	23.85

It can be deduced that somewhere around 15%-30% of the whole data being It2Ro might provide good results. Also, it can be remarked that the proposed model performs better on mixed data with an intuitively very strong diversity characteristic.

3.5 Final Experiments

We continued by using RSC, Vox Populi, RTASC, RSS and CoRoLa on top of the previous CV (all datasets described as previously noted in Table 1) and from the It2Ro CV dataset data from all the splits (including test). The models were trained for 10 epochs, and the obtained performance can be deduced from Table 10.

Table 10. Best results during the all-data experiment.

It2Ro	LR	Epoch	CV_Ro		ALL1		ALL2		USPDATRO	
			WER %	CER %	WER %	CER %	WER %	CER %	WER %	CER %
-	1e-6	3	18.65	6.18	20.61	9.19	14.93	7.42	37.33	18.80
15k	1e-6	2	21.82	6.75	21.58	9.29	17.49	7.98	38.08	18.24
25k	1e-6	3	20.82	6.89	21.87	9.84	14.15	6.85	39.68	19.47
55k	1e-6	7	20.54	6.55	21.54	9.87	13.10	6.00	36.57	17.57
75k	1e-6	3	21.20	21.64	21.64	9.65	15.68	7.53	36.61	18.31

It can be seen that on some configurations the model keeps performing better on ALL2 and USPDATRO datasets as in the initial experiment, while it continues to sit below on the other datasets adding ALL1. The reason behind it can be interpreted from a lexical perspective, since the Italian audio provides some characteristics which aren't balanced in any way in order for the model to remain focused on a lexical correct Romanian.

4 Conclusion and Future Work

The presented experiments show the improvement of training the Whisper model with additional Italian speech seen during evaluation on mixed data with an increased diversity characteristic, while keeping close to the metrics on other datasets. The proposed method can be also applied to other language families where there is a difference in the quantity of resources, or where a better accuracy is needed with less computational power. Regarding real-life applicability, systems that want to ensure performance in disregard to the speaker's demographic information can benefit from such inclusion of data in the development process.

Future work proposes to include more Romance languages into the data conversion process because they might bring considerable value to the spectrum of voices and accents and might also calibrate the loss in metrics for some of the datasets. Also, better phoneme mappings are intended with more extensive phoneme sets in the hopes of increasing the quality of the newly composed data and the overall performance of the model.

References

- Ahn, E., & Chodroff, E. 2022. VoxCommunis Corpus. *OSF*. <https://osf.io/t957v>
- Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., Engel, J., Fang, W., Fan, L., Fougner, C., Gao, L., Gong, C., Hannun, A., Han, T., Johannes, L. V., Jiang, B., Ju, C., Jun, B., LeGresley, P., Lin, L., Liu, J., Liu, Y., Li, W., Li, X., Ma, D., Narang, S., Ng,

- A., Ozair, S., Peng, Y., Prenger, R., Qian, S., Quan, Z., Raiman, J., Rao, V., Satheesh, S., Seetapun, D., Sengupta, S., Srinet, K., Sriram, A., Tang, H., Tang, L., Wang, C., Wang, J., Wang, K., Wang, Y., Wang, Z., Wang, Z., Wu, S., Wei, L., Xiao, B., Xie, W., Xie, Y., Yogatama, D., Yuan, B., Zhan, J., and Zhu, Z. 2016. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, 48:173-182. New York, NY, USA: JMLR.org.
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. 2020. Common Voice: A Massively - Multilingual Speech Corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211-4215.
- Avram, A.-M., Păiș, V. F., and Tufiș, D. 2020. Towards a Romanian End-to-End Automatic Speech Recognition Based on Deepspeech2. In *Proceedings of the Romanian Academy Series A*. 21:395-402.
- Avram, A.-M., Păiș, V. F., and Tufiș, D. 2021. Romanian Speech Recognition Experiments from the ROBIN Project. In *The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR)*, pages 103-114.
- Avram, A.-M., Smădu, R.-A., Păiș, V., Cercel, D.-C., Ion, R., and Tufiș, D. 2023. Towards Improving the Performance of Pre-Trained Speech Models for Low-Resource Languages Through Lateral Inhibition. In *2023 46th International Conference on Telecommunications and Signal Processing (TSP)*, pages 234-237. Prague, Czech Republic. DOI: 10.1109/TSP59544.2023.10197791.
- Baevski, A., Zhou, H., Mohamed, A., and Auli, M. 2020. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 12449-12460. Vancouver, BC, Canada: Curran Associates Inc.
- Barbu-Mititelu, V., Irimia, E., and Tufiș, D. 2014. CoRoLa - The Reference Corpus of Contemporary Romanian Language. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1235-1239. Reykjavik, Iceland: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2014/pdf/360_Paper.pdf.
- Cerniavski, R., & Stymne, S. 2023. Multilingual Automatic Speech Recognition for Scandinavian Languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 460-466. University of Tartu Library. <https://aclanthology.org/2023.nodalida-1.46>.
- Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa J., Rivera, C., and Bapna, A. 2022. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798-805. <https://api.semanticscholar.org/CorpusID:249062909>.
- Cristea, D., Pistol, I., Boghiu, Ș., Bibiri, A.-D., Gîfu, D., Scutelnicu, A., Onofrei, M., Trandabăț, D., and Bugeag G. 2020. CoBiLiRo: A Research Platform for Bimodal Corpora. In *Proceedings of the 1st International Workshop on Language Technology*

- Platforms*, pages 22-27. Marseille, France: European Language Resources Association (ELRA). <https://aclanthology.org/2020.iwlt-p-1.4>.
- Domokos, J., Buza, O., and Todorean, G. (2014, 06). Romanian Phonetic Transcription Dictionary for Speeding up Language Technology Development. In *Language Resources and Evaluation*, 49:311-325. DOI: 10.1007/s10579-013-9262-z.
- Georgescu, A.-L., Cucu, H., and Burileanu, C. 2021. Improvements of Speed's Romanian ASR System During ReTeRom Project. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 177-182. DOI: 10.1109/SpED53181.2021.9587383.
- Georgescu, A.-L., Cucu, H., Buzo, A., and Burileanu, C. 2020. RSC: A Romanian Read Speech Corpus for Automatic Speech Recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6606-6612. Marseille, France: European Language Resources Association. Accessed and retrieved from <https://www.aclweb.org/anthology/2020.lrec-1.814>.
- Hauer, B., Habibi, A. A., Luan, Y., Mallik, A., and Kondrak, G. 2020. Low-Resource G2P and P2G Conversion with Synthetic Training Data. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117-122. Association for Computational Linguistics. DOI: 10.18653/v1/2020.sigmorphon-1.12.
- Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., and Gorman, K. 2020. Massively Multilingual Pronunciation Modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223-4228. Marseille, France: European Language Resources Association. <https://aclanthology.org/2020.lrec-1.521>.
- Lőrincz, B., Irímia, E., Stan, A., and Barbu Mititelu, V. (2023). RoLEX: The Development of an Extended Romanian Lexical Dataset and Its Evaluation at Predicting Concurrent Lexical Information. In *Natural Language Engineering*, 29(3):720-745. DOI: 10.1017/S1351324922000419.
- Păiș, V., Ion, R., Avram, A.-M., Irímia, E., Mititelu, V. B., and Mitrofan, M. 2021. Human-Machine Interaction Speech Corpus from the ROBIN project. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 91-96. IEEE. DOI: 10.1109/sped53181.2021.9587355.
- Păiș, V., Irímia, E., Barbu-Mititelu, V., and Ion, R. 2023. USPDATRO: Underrepresented Speech Dataset from Romanian language Open Data. *Zenodo*. DOI: 10.5281/zenodo.7898232.
- Păiș, V., Mititelu, V.B., Ion, R., and Irímia, E. 2023. Evaluating a Fine-Tuned Whisper Model on Underrepresented Romanian Speech. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 141-145.
- Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. 2015. Librispeech: An ASR Corpus Based on Public Domain Audio Books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206-5210.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the*

- 40th International Conference on Machine Learning, 202:28492-28518. PMLR. <https://proceedings.mlr.press/v202/radford23a.html>.
- Stan, A. 2020. RECOApy: Data Recording, Pre-Processing and Phonetic Transcription for End-to-End Speech-Based Applications. In *Proceedings of Interspeech 2020*, pages 586-590. DOI: 10.21437/Interspeech.2020-1184.
- Stan, A., Dinescu, F., Tiple, C., Meza, S., Orza, B., Chirila, M., and Giurgiu, M. 2017. The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset. In *Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD 2017)*, pages 1-6. Bucharest, Romania. DOI: 10.1109/SPED.2017.7990428.
- Stan, A., Lőrincz, B., Nuțu, M., and Giurgiu, M. 2021. The MARA corpus: Expressivity in End-to-End TTS Systems Using Synthesised Speech Data. In *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 85-90. DOI: 10.1109/SpED53181.2021.9587438.
- Stan, A., Yamagishi, J., and Aylett, S. K. 2011. The Romanian Speech Synthesis (RSS) Corpus: Building a High Quality HMM-Based Speech Synthesis System Using a High Sampling Rate. In *Speech Communication*, 53(3):442-450. Accessed and retrieved from <https://doi.org/10.1016/j.specom.2010.12.002>.
- Strubell, E., Ganesh, A., and McCallum, A. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645-3650. Florence, Italy: Association for Computational Linguistics. DOI: 10.18653/v1/P19-1355.
- Toma, S.-A., Stan, A., Pura, M., and Barsan, T. 2017. MaRePhoR — An Open Access Machine-Readable Phonetic Dictionary for Romanian. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 1-6. Bucharest, Romania. DOI: 10.1109/SPED.2017.7990435.
- Wang, C., Riviere, M. a., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, Volume 1: Long Papers, pages 993-1003. Association for Computational Linguistics. <https://aclanthology.org/2021.acl-long.80>.
- Zgank, A. 2019. Cross-Lingual Speech Recognition between Languages from the Same Language Family. In *Romanian Academy Series A*, 20:184-191.

INDEX OF AUTHORS

- Alexei Victoria: 101
Badea Valentin: 85
Barbu Mititelu Verginica: 85, 165
Bejinariu Silviu-Ioan: 199
Bobicev Victoria: 73, 101
Botoșineanu Luminița: 199
Bumbu Tudor: 73
Caftanатов Olesea: 63,73
Caganovschi Daniela: 63
Chan Lun Min: 135
Ciocoiu Călina-Georgiana: 187
Cohal Alexandru Laurențiu: 199
Cristea Dan: 23
Dinu Liviu: 18
Florea Bogdan: 119
Gasar Carol-Luca: 235
Gifu Daniela: 177
Haja Gabriela: 15, 23
Hlaing Moe Aung: 135
Iftene Adrian: 119, 215
Ion Radu: 17, 85, 149
Ionescu Radu Tudor: 19
Irimia Elena: 85
Khant Min Kaung: 135
Luca Ramona: 199
Malahov Ludmila: 63, 73
Mincu Eugenia: 63
Myint Swe Bhone: 135
Nastasiu Ana: 73
Nechifor Andreea: 1
Olariu Florin-Teodor: 199
Olariu Veronica: 199
Parahonco Alexandr: 43
Păiș Vasile: 85, 235
Petic Mircea: 43
Popescu Marius: 16
Sann Thit Aung: 135
Sterea-Grossu Cosmin: 85
Stoica Dan S.: 177
Su Phyoe Aye: 135
Tamba Elena Isabelle: 23
Tălămbuță Dan: 63, 73
Thaw Phyoe Min: 135
Thu Htet Phyoe: 135
Tufiș Dan: 23, 85
Țițchiev Inga: 63, 73
Vasilița Irina: 215
Voicu Tudor: 165