

PROCEEDINGS OF THE 16th
INTERNATIONAL CONFERENCE
“LINGUISTIC RESOURCES AND TOOLS
FOR NATURAL LANGUAGE PROCESSING”,
ONLINE, 13-14 DECEMBER 2021

Editors

Petru Rebeja
Mihaela Onofrei
Dan Cristea
Dan Tufiş

Organisers

Institute for Computer Science
Romanian Academy, Iaşi Branch

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

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

Romanian Association of Computational Linguistics

Under the auspices of the Academy of Technical Sciences of Romania

The publication of this volume was supported by
the Faculty of Computer Science,
“Alexandru Ioan Cuza” University of Iași

ISSN 1843-911X

SCIENTIFIC COMMITTEE

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

Anca Bibiri, Department of Social Science and Humanities, Institute of Interdisciplinary Research, “Alexandru Ioan Cuza” University of Iași

Tiberiu Boroș, Adobe Bucharest

Alexandru Ceașu, European Commission

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

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

Horia Cucu, Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest

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

Nils Diewald, Leibniz-Institut für Deutsche Sprache, Mannheim, Germany

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

Daniela Gifu, Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași and Institute for Computer Science, Romanian Academy, Iași Branch

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

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, Canada

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

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

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

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

Mihaela Onofrei, Institute for Computer Science, Romanian Academy, Iași Branch and Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

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

Adriana Ștan, Communications Department, Faculty of Electronics, Telecommunications and Information Technology, Technical University of Cluj-Napoca

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

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

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

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

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

ORGANISING COMMITTEE

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

Paula Alexandra Crucianu, Institute for Computer Science, Romanian Academy, Iași Branch

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

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

Daniela Gifu, Institute for Computer Science, Romanian Academy, Iași Branch and Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

Gabriela Haja, “Alexandru. Philippide” Institute of Romanian Philology, Romanian Academy, Iași Branch

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

Mihaela Onofrei, Institute for Computer Science, Romanian Academy, Iași Branch and Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași

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

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

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

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

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

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

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

TABLE OF CONTENTS

FOREWORD.....	VII
CONFERENCE KEYNOTE SPEECHES	1
FROM RESEARCH TO APPLIED USE CASES: HOW RECENT DEVELOPMENTS IN NLP AND SPEECH ARE CHANGING AI SERVICES	1
<i>Dmitry Turchyn</i>	
TOWARDS DIGITAL LANGUAGE EQUALITY IN 2030 – CONTRIBUTIONS FROM EUROPEAN LANGUAGE GRID AND EUROPEAN LANGUAGE EQUALITY	1
<i>Georg Rehm</i>	
PROVIDING STAGED CORPUS ACCESS AT MULTIPLE LEVELS TO MAXIMIZE CORPUS USABILITY AND MINIMIZE COSTS.....	2
<i>Marc Kupietz</i>	
LANGUAGE TECHNOLOGIES AND LARGE-SCALE DIGITAL HERITAGE RESOURCES: FRIENDS OR FOES?	2
<i>Milena Dobрева</i>	
DELORO WORKSHOP KEYNOTE SPEECHES	3
DUOSEARCH – A NOVEL SEARCH ENGINE FOR BULGARIAN HISTORICAL DOCUMENTS.....	3
<i>Ivan Koychev, Angel Beshirov, and Suzan Hadzhieva</i>	
ACCESSING GREEK HISTORICAL HANDWRITTEN DOCUMENTS USING THE MDOC.TS PLATFORM.....	3
<i>Ioannis Pratikakis</i>	
CHAPTER 1. ROMANIAN LINGUISTIC RESOURCES.....	5
LINKED OPEN DATA RESOURCES FOR ROMANIAN	7
<i>Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Maria Mitrofan, Andrei- Marius Avram, and Eric Curea</i>	
DEVELOPING LINGUISTIC RESOURCES FOR ROMANIAN WRITTEN AND SPOKEN LANGUAGE.....	21
<i>Oana Niculescu</i>	
DIGITAL RESOURCES DEVELOPED WITHIN THE BIBLIO-MLROM PROJECT	37
<i>Cosmina-Maria Berindei</i>	
CHAPTER 2. CURATION OF DIGITAL HERITAGE.....	47
THE ROLE OF THE PHILOLOGIST IN THE PROJECT ARTIFICIAL INTELLIGENCE MODELS (DEEP LEARNING) APPLIED IN THE ANALYSIS OF OLD ROMANIAN LANGUAGE (DELORO)	49
<i>Gabriela Haja and Elena Isabelle Tamba</i>	

DATA STRUCTURE AND ACQUISITION IN DeLORO – A TECHNOLOGY FOR DECIPHERING OLD CYRILLIC-ROMANIAN DOCUMENTS	59
<i>Dan Cristea, Cristian Pădurariu, Petru Rebeja, Andrei Scutelnicu, and Mihaela Onofrei</i>	
PRESERVATION OF ROMANIAN LINGUISTIC HERITAGE. FRAMEWORK FOR DIALECTAL DATA MANAGEMENT	75
<i>Silviu-Ioan Bejinariu, Florin Iftene, Manuela Nevaci, and Carmen Irina Floarea</i>	
OLD ROMANIAN LEXICONS AND THEIR REPRESENTATION IN DLR	93
<i>Mihai Alex Moruz and Mădălina Ungureanu</i>	
CHAPTER 3. DATASETS AND CORPORA.....	103
RoITD: ROMANIAN IT QUESTION ANSWERING DATASET.....	105
<i>Dragoş Constantin Nicolae and Dan Tufiş</i>	
A PUZZLE-BASED DATASET FOR NATURAL LANGUAGE INFERENCE	119
<i>Roxana Szomiu and Adrian Groza</i>	
BUILDING A CORPUS OF MEDICAL PARAPHRASES IN ROMANIAN	139
<i>Ioana Buhnă</i>	
CHAPTER 4. SPEECH AND TEXT MODELS FOR ROMANIAN LANGUAGE.....	153
A NEW METHOD OF EMPHASIZING THE FORMANTS OF THE VOCAL SPECTRUM	155
<i>Marius Zbancioc and Monica Feraru</i>	
ROMANIAN TEXT ANONYMIZATION EXPERIMENTS FROM THE CURLICAT PROJECT	165
<i>Vasile Păiş, Elena Irimia, Radu Ion, Dan Tufiş, Maria Mitrofan, Verginica Barbu Mititelu, Andrei-Marius Avram, and Eric Curea</i>	
CHAPTER 5. LANGUAGE VARIETIES AND SOCIAL MEDIA	179
STATISTICS OF SOME REGIONAL VARIETIES OF ROMANIAN GREETINGS	181
<i>Anca-Diana Bibiri, Mihaela Colhon, and Mihaela Mocanu</i>	
OFFENSIVE LANGUAGE IDENTIFICATION IN SOCIAL MEDIA FOR ROMANIAN LANGUAGE	199
<i>Mihaela Onofrei, Diana Trandabăţ, and Bogdan-Andrei Donu</i>	
A SURVEY ON FAKE NEWS DETECTION TECHNIQUES.....	207
<i>Radu Ciora, Marius Cioca, and Daniela Gifu</i>	
INDEX OF AUTHORS.....	215

FOREWORD

This volume includes the papers of the 16th edition of ConsILR (traditional acronym for Consorțiul de Informatizare pentru Limba Română – Consortium of Informatization for Romanian Language), the International Conference on Linguistic Resources and Tools for Natural Language Processing, held between 13-14 December 2021, together with the first Workshop of the project “DeLORo – Deep Learning for Old Romanian”. The scientific events were organized by two institutes of the Romanian Academy, the Institute of Computer Science in Iași and the Institute of Artificial Intelligence "Mihai Drăgănescu" in Bucharest, together with the Faculty of Computer Science of the "Alexandru Ioan Cuza” University of Iași and the Romanian Association of Computational Linguistics. As in most of the previous editions, the event ran under the auspices of the Technical Sciences Academy of Romania.

More than once, in the past, the presentations were lively broadcast on the web during the events, but for the last two editions, given the restrictions imposed in the context of Covid-19 pandemic, the communications were entirely online. We are grateful to all the virtual attendants of the 16th ConsILR Conference, to the reviewers who ensured a quality selection of the papers and to the organizers who skilfully managed the online event. From its first edition, in 2001, the ConsILR Conference 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 natural languages, with a special emphasis on the Romanian, from a computational perspective. The series of events have run, with few exceptions, every year, first in the format of a workshop, and since 2010 as an international conference.

The conference program, mirrored in this volume, was dense, with 15 presentations of the latest results of researchers in this field and four keynote lectures delivered by reputed specialists in the area of NLP-centred AI. Two other keynote speeches were hosted within the DeLORo workshop. Abstracts of these lectures are included at the beginning of the volume.

The contributed articles were organized in five sections: 1. Romanian Linguistic Resources, 2. Curation of Digital Heritage, 3. Datasets and Corpora, 4. Speech and Text Models for Romanian Language, and 5. Language Varieties and Social Media.

The topics addressed are of real interest, from linked linguistic open data, resources developed in projects, old language curation, and corpora to standardized methods of representation of linguistic resources, speech processing topics, applications for recognition of false news and offensive language. Most of the results presented at the conference are public, open to anyone interested in taking them over and using them.

December 2021

The editors

CONFERENCE KEYNOTE SPEECHES

From Research to Applied Use Cases: How Recent Developments in NLP and Speech are Changing AI Services

DMITRY TURCHYN

Microsoft

dturchyn@microsoft.com

In this session, you will learn how the new generation of ML models is changing what's possible with AI, and how it transforms into applied services that can be used by organizations to transform their products and processes. You'll see a demo of some of the state-of-the-art AI services, as well as learn about practical use cases being solved with Speech and Language AI services, based on examples of challenges solved by organizations in our region.

Towards Digital Language Equality in 2030 – Contributions from European Language Grid and European Language Equality

GEORG REHM

German Research Center for Artificial Intelligence

georg.rehm@dfki.de

Europe is a multilingual society, in which dozens of languages are spoken. The only option to enable and to benefit from multilingualism is through Language Technologies (LT), i.e., Natural Language Processing and Speech Technologies. The presentation will provide an overview of the European Language Grid (ELG), which is targeted to evolve into the primary platform and marketplace for LT in Europe by providing one umbrella platform for the European LT landscape, including research and industry, enabling all stakeholders to upload, share and distribute their services, products and resources. At the end of our EU project, which will establish a legal entity in 2022, the ELG will provide access to more than 750 services for all European languages as well as thousands of data sets; already now the ELG platform contains more than 5000 resources. Furthermore, the presentation will provide an overview of ELG's sister project, European Language Equality (ELE) as well as preliminary results. This EU project develops

a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

Providing Staged Corpus Access at Multiple Levels to Maximize Corpus Usability and Minimize Costs

MARC KUPIETZ

Leibniz Institute for the German Language

kupietz@ids-mannheim.de

Big Corpus data are notoriously difficult to use linguistically because of their high dimensionality, their opaque structure and because of IPR and license restrictions. In my talk, I will introduce a multi-level approach used by the corpus linguistics group at IDS Mannheim in the context of KorAP, DeReKo and the European Reference Corpus initiative EuReCo, to make corpus as usable as possible, at feasible costs and despite these challenges. I will use examples from the Reference Corpus of the Contemporary Romanian Language (CoRoLa) and report on preliminary results.

Language Technologies and Large-Scale Digital Heritage Resources: Friends or Foes?

MILENA DOBREVA

UCL Qatar

milena.dobрева@gmail.com

The talk will explore the relationship between large scale digital heritage resources and language technologies. It will discuss the different roles and expectations of researchers – end-users of digital resources, and the creators of these resources and will be focused mostly on the applications of data science for innovative research in digital cultural heritage collections and the emerging innovation labs in libraries. The talk will provide examples from the use of digitised collections from Europeana, the Library of Congress, the British Library, the Royal Library of Denmark, Miguel de Cervantes Virtual Library, and the National Library Ivan Vazov in Plovdiv. Finally, the talk will also look into the expectations and challenges around the much-discussed application of artificial technologies in the domain of digital cultural heritage.

DELORO WORKSHOP KEYNOTE SPEECHES

DuoSearch – A Novel Search Engine for Bulgarian Historical Documents

IVAN KOYCHEV, ANGEL BESHIROV, AND SUZAN HADZHIEVA

“St Kliment Ohridski” University of Sofia

koychev@fmi.uni-sofia.bg, angel.beshirov@abv.bg,
suzanhadzhieva@gmail.com

Search in collections of digitized historical documents is hindered by a two-prong problem, orthographic variety, and optical character recognition (OCR) mistakes. We present a new search engine for historical documents, DuoSearch, which uses ElasticSearch and machine learning methods based on deep neural networks to offer a solution to this problem. It was tested on a collection of historical newspapers in Bulgarian from the mid-19th to the mid-20th century. The system provides an interactive and intuitive interface for the end-users, allowing them to enter search terms in modern Bulgarian and search across historical spellings. This is the first solution facilitating the use of digitized historical documents in Bulgarian.

Accessing Greek Historical Handwritten Documents Using the μ Doc.tS Platform

IOANNIS PRATIKAKIS

Democritus University of Thrace

ipratika@ee.duth.gr

The potential to access our written past, which stimulates the interest of not only researchers but also the general public, makes Handwritten Text Recognition (HTR) and KeyWord Spotting (KWS) a highly appealing set of technologies, among the ones appearing in the document image analysis research area. This talk will strive towards a manifestation for the presentation of the HTR and KWS technologies that have been developed in the context of the research project ‘A platform for the transcription of historical handwritten documents – μ Doc.tS’. The performance of these technologies will be exemplified for a set of handwritten document collections originating from the Stavronikita Monastery on Mount Athos that recently have been made publicly available for research purposes.

CHAPTER 1

ROMANIAN LINGUISTIC RESOURCES

Linked Open Data Resources for Romanian

VERGINICA BARBU MITITELU, ELENA IRIMIA, VASILE PĂIȘ,
MARIA MITROFAN, ANDREI-MARIUS AVRAM,
and ERIC CUREA

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

{vergi, elena, vasile, maria, andrei.avram,
eric}@racai.ro

Abstract

In this paper we present several new Linked Open Data resources for Romanian: the treebanks SiMoNERo, PARSEME-Ro and LegalNERo, and the lexicon RoLEX. These resources were initially developed in other formats, but they have been converted to observe the Linked Data formalism. They were published on a dedicated website, while querying them using SPARQL is also provided on an Apache Jena Fuseki server. Their metadata have been uploaded into the Linguistic Linked Open Data Cloud to increase their visibility and accessibility.

1 Introduction

In the context of a growing number of resources becoming available worldwide as allowed by the new technologies and by the increased awareness of the importance of releasing publicly relevant datasets, there is a need for ensuring their interoperability and integration with each other: resources must have specific formats, must follow certain communication protocols so as to ensure exchange of data (i.e., they must be syntactically interoperable) and must “automatically interpret exchanged information meaningfully and accurately” [Ide and Pustejovsky, 2010] (that is, they must be semantically interoperable). These requirements are imposed at several levels: metadata, data categories, publication format and software sharing. The set of best practices in publishing structured data on the Web is at the core of the *linked data* concept. A license that allows free access to the resource ensures their *open* characteristic. In this paper we discuss only about resources from the domain of linguistics, so the term to cover them will be Linguistic Linked Open Data (LLOD henceforth).

This paper presents several resources for the Romanian language that are now in Linked Data format (Section 2) and have been made freely available in specific ways (see Section 3). Besides being interlinked with well-known ontologies, some of them have also been interlinked with each other (see Section 4) in an effort to make information easily accessible from one resource to another. In Barbu Mititelu *et al.* [2020], we presented the LLOD format for the Romanian wordnet

(RoWN) [Tufiş and Barbu Mititelu, 2014], of the Romanian Reference Treebank (RRT) [Barbu Mititelu, 2018], and of corpus-driven linguistic data, more exactly word frequencies, lemma frequencies and word embeddings extracted from the Corpus of Contemporary Romanian Language (CoRoLa) [Tufiş *et al.*, 2019]. In this paper we present the newly available Romanian LLOD resources, while the discussion about their publishing (Section 3) and interlinking (Section 4) will refer to all resources we have created (both those reported in [Barbu Mititelu *et al.*, 2020] and the newly converted ones).

2 Resources

Two types of language resources are presented here: treebanks and lexicons. The former is instantiated by three Romanian treebanks, one of which is a new development (see Subsection 2.1), while the others are older, but newly converted to the LD format (see Subsections 2.2 and 2.3). Only one lexicon is presented below (see Subsection 2.4), which is a recent development.

2.1 LegalNERo

LegalNERo¹ is a gold annotated corpus for named entity recognition (NER) in the Romanian legal domain. It is extracted from the larger MARCELL-RO² corpus [Tufiş *et al.*, 2020], enriched with manual annotations. The corpus is tokenized, part-of-speech tagged, lemmatized and syntactically parsed with UDPipe³ [Straka *et al.*, 2016]. It contains 8,248 sentences in which 265,335 tokens occur.

The main reason behind LegalNERo’s creation was improving the performance of the existing Romanian NER system [Păiş, 2019] in the legal domain. Considering the specifics associated with the legal domain, we employed both general-domain entity classes (person names, organizations, locations, time expressions) and a domain-specific class (legal reference). When introducing the legal reference class, we considered it to be similar to the work of [Landthaler *et al.*, 2016] and the coarse-grained class proposed by [Leitner *et al.*, 2019], without additional sub-classes.

The annotation of the LegalNERo corpus was performed by 5 human annotators, supervised by two senior researchers, using the BRAT tool⁴ [Stenetorp *et al.*, 2012], integrated in the RELATE platform [Păiş *et al.*, 2020]. In order to allow the corpus to be used for training different NER systems, we transformed

¹ <https://doi.org/10.5281/zenodo.4772094>

² The MARCELL-RO corpus was developed within the “Multilingual Resources for CEF.AT in the legal domain” (MARCELL) project, in which a large comparable corpus [Váradi *et al.*, 2020] was produced comprising national legislation in 7 European languages, including Romanian. See project website: <https://marcell-project.eu/>

³ <https://ufal.mff.cuni.cz/udpipe>

⁴ <https://brat.nlplab.org/>

the span-based annotations to token-based annotations. For tokenization purposes we used UDPipe [Straka and Straková, 2017]. The legal reference entities were annotated with sub-entities, considering organizations and time expressions. This process produced 4 different views over the same corpus: a) span-based considering all entities and sub-entities; b) span-based without the legal reference class, but considering the sub-entities (without overlapping spans); c) token-based with all classes, without sub-entities; d) token-based without the legal reference, considering the sub-entities. Finally, we linked location entities to the GeoNames⁵ ontology (whenever possible).

Table 1 Span-based named entities statistics

Dataset	LEGAL	PER	LOC	ORG	TIME	GEO	TOTAL NEs
Without Legal Ref	-	914	2,276	6,209	4,643	-	14,042
Without sub-entities	3,387	914	2,276	4,824	2,213	-	13,614
All annotations	3,387	914	2,276	6,209	4,643	-	17,429

The presence of multiple perspectives over the corpus and the linking to an ontology made it natural to convert the LegalNERo corpus into a LLOD format. After conversion, we obtained 5,761,781 triples, describing the different levels of the corpus: span-based, token-based, GeoNames linking. Additional statistics, considering the annotated text-spans can be found in Table 1.

The corpus was exploited to create a NER system using neural network architectures with pre-trained word embeddings. Different NER experiments are presented in [Păiș and Mitrofan, 2021] and [Păiș *et al.*, 2021].

2.2 SiMoNERo

SiMoNERo [Barbu Mititelu and Mitrofan, 2020] is a Romanian medical treebank, with texts from cardiology, endocrinology and diabetes. The corpus is segmented at the sentence level, morphologically annotated and lemmatized, using the TTL tool [Ion, 2007]. For syntactic parsing, the NLP-Cube analyzer [Boros *et al.*, 2018] trained on RRT was used. Medical entities were manually annotated by two annotators using four semantic groups from UMLS⁶ (i.e., a thesaurus and ontology of biomedical concepts): anatomical parts, chemicals and drugs, disorders and medical procedures. Their frequencies in the corpus are rendered in Table 2.

⁵ <https://www.geonames.org/>

⁶ <https://www.nlm.nih.gov/research/umls>

SiMoNERo contains 4,681 phrases and 146,020 tokens. It is a gold resource for the morphological level and also for the medical named entities annotation.

Table 2 Types of medical entities in SiMoNERo: ANAT = anatomical parts, CHEM = chemicals and drugs, DISO = disorders and PROC = medical procedures

ANAT	CHEM	DISO	PROC	TOTAL
1,964	4,156	6,611	1,402	14,133

For the exploitation of this corpus, we converted it from its initial CONLL-U format into BRAT format; thus, we could train a NER system which uses a recurrent neural network, based on BiLSTM cells, and a final layer based on CRF. The word embeddings used were calculated on a biomedical corpus. The preliminary results obtained showed an improvement of the F1-score for the NER task of 0.5 compared with [Mitrofan, 2019].

Table 3 Types of verbal multiword expressions in PARSEME-Ro: VID = verbal idioms, LVC.full = light verb constructions, LVC.cause = causal light verb constructions, IRV = inherently reflexive verbs

VID	LVC.full	LVC.cause	IRV	TOTAL
1,611	313	183	3,784	5,891

2.3 PARSEME-Ro

Aware of the challenges expressions pose in the process of natural language processing [Sag *et al.*, 2002], our group actively participated in the PARSEME COST Action⁷ and in the activities of classification, description and guidelines for annotating expressions⁸, for creating linguistic resources [Barbu Mititelu *et al.*, 2019] and tools [Boros *et al.*, 2017] for identifying expressions in texts. The corpus we created in this Action and which we call PARSEME-Ro is annotated morphologically, syntactically and with verbal multiword expressions: it contains journalistic texts, counting 56,703 sentences and 1,015,623 tokens. We manually annotated 5,891 verbal multiword expressions of four types (see Table 3) according to the instructions agreed within PARSEME and checked for their consistency [Ramisch *et al.*, 2018]. The morphologic and syntactic annotation was done automatically, using UDPipe trained on RRT.

The corpus has been exploited in all editions of the PARSEME shared tasks so far [Savary *et al.*, 2018; Ramisch *et al.*, 2018; Ramisch *et al.*, 2020] for training,

⁷ <https://typo.uni-konstanz.de/parseme/>

⁸ <https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=home>

developing and testing the systems participating in the competition for identifying verbal multiword expressions.

Both SiMoNERo and PARSEME-Ro treebanks were converted to LLOD specifications using the CoNLL-U to RDF tool from the Applied Computational Linguistics Lab [Chiarcos and Fäth, 2017]. The tool converted the files in a Turtle format and, subsequently, into XML/RDF by means of a Python script⁹. As reported for RRT [Barbu Mititelu *et al.*, 2020], this conversion variant was preferred so as to preserve the order of sentences in the treebanks files. The LLOD SiMoNERo contains 1,396,182 triples, while the LLOD PARSEME-Ro corpus contains 11,203,698 triples.

2.4 RoLEX

RoLEX phonological lexicon, developed in the ReTeRom project¹⁰, is the largest validated resource of this type available for Romanian (330,866 entries). It was created based on the vocabulary of the textual component of a speech corpus that contained data from the Romanian Wikipedia, news, interviews, talkshows, spontaneous speech, fairy tales, novels, etc. Morphosyntactic and lemma information in RoLEX was retrieved from a large in-house general lexicon for Romanian. Syllabification, stress and phonetic transcription information is based, partially, on previous resources like RoSyllabiDict [Barbu, 2008] and MaRePhor [Toma *et al.*, 2017] and otherwise automatically generated using Stan *et al.*'s [2011] tool. The aggregation of all this information went through a systematic correction process, either manual or automatic (when rules could be implemented). In its initial, tabular, format, RoLEX entries have 6 columns that contain information for: the represented form, the lemma associated with that form, the morphosyntactic description (MSD), syllabification, stress and phonetic transcription.

For its conversion to LLOD specification, we used the following classes and attributes (most of them part of the lexicon model for ontologies (*lemon*) developed by the Ontology-Lexica (Ontolex¹¹) community group):

- The unique lemmas are encoded using the main lemon class, *LexicalEntry*, which is the basic unit of the lexicon, associated with a part of speech, a morphological pattern, an etymology and a set of senses. RoLEX does not include etymological information, while the morphological pattern is replaced by a list of lemon Form objects that represent the set of forms (i.e., the inflectional paradigm) associated to that lemma. The set of senses is not explicitly encoded in RoLEX, but the interlinking between RoLEX and

⁹ https://github.com/racai-ai/RoLLOD/tree/master/conllu_convertors

¹⁰ <https://www.racai.ro/p/reterom/>

¹¹ <https://www.w3.org/2016/05/ontolex/#overview>

RoWN compensates for this absence (see Section 4.2). Here is an example of a lexical entry for the lemma *copil* “child”:

```
:lex_copil_n a ontolex:LexicalEntry;
rdfs:label "copil_n"@ro;
ontolex:canonicalForm :form_copil_n;
wn:partOfSpeech wn:n;
ontolex:lexicalForm:form_copil_n_noun_ind_masc_plur;
ontolex:lexicalForm:form_copil_n_noun_acc_nom_def_masc_plur;
ontolex:lexicalForm:form_copil_n_noun_dat_gen_def_masc_plur;
etc.
```

- `ontolex:canonicalForm` property links the lexical entry to (at most) a canonical form (i.e., the lemma);
- `rdfs:label` property is also recommended by Ontolex for redundantly encoding the canonical form, to assure compatibility with RDFS-based representation systems;
- `wn:partOfSpeech` is used to separately encode the part of speech of the lexical entry;
- `ontolex:lexicalForm` is used to enumerate a list of morphological forms associated with the canonical form;
- Each object in the list of lexical forms is described using `ontolex:writtenRep`, `ontolex:phoneticRep` and `conll:POS` properties. `PhoneticRep` encodes a single phonetic representation of the described form, using Romanian version of Speech Assessment Methods Phonetic Alphabet (SAMPA¹²). In the absence of specific lemm properties to encode syllabification and stress information, we used `writtenRep` with two dedicated labels, `@syl` and `@stress`. The graphical representation of the form is also encoded with `writtenRep`, with an associated label `@ro`. `Conll:POS` describes the MSD label of the form. Here is the description for the lexical form *copilul* “the child”:

```
:form_copil_n_noun_acc_nom_def_masc_sing a ontolex:Form;
ontolex:writtenRep "copilul"@ro;
conll:POS "Ncmsry";
ontolex:writtenRep "co.pi.lul"@syl;
ontolex:writtenRep "cop'ilul"@stress;
ontolex:phoneticRep "k o p i l u l"@ro-RO-sampa .
```

¹² <https://www.phon.ucl.ac.uk/home/sampa/>

The unique identifiers of the lexicalForm objects were generated by automatically expanding the MSD labels into UD¹³ feature lists. For each part of speech, we manually designed feature mapping sheets, based on the UD feature descriptions¹⁴, but extending them when the information encoded in the MSD had no correspondence in the UD model: e.g., we used a Polarity=Pos/Neg property for conjunctions; we created a psor property to specify the possessor number for possessive pronouns and determiners; we created a property to specify the pre/post-nominal position of the demonstrative pronouns: Position=Prenom/Postnom. We used the feature mapping sheets to automatically generate a feature list for each MSD and designed form labels by concatenating the lemma, the part of speech (POS) label, the UD part of speech (UPOS) label and the features in alphabetical order. The necessity to encode both the POS and UPOS labels comes from their partial overlap: e.g., in the MSD convention, all nouns have the same part of speech, while the UD convention distinguishes between proper and common nouns (PROPN vs. NOUN labels) at UPOS level.

There are, in Romanian, different types of ambiguities that concern the phonetic transcription of a graphical form, thus posing challenges for automatic speech recognition systems; some of them manifest at the stress or syllabification level:

- There are graphically identical inflected forms of a specific lemma that are differently stressed: e.g., the present form 'află versus the past simple form afl'ă ; however, this does not result in an ambiguity problem in RoLEX, since each inflected form has a specific entry with an associated phoneticRep property;
- There is also ambiguity at the lemma level, when:
 - Different stressed forms exist for the same word, both accepted by the standard norms of the language: e.g., *antic* “ancient” can be stressed as either 'antic or ant'ic; for this case, alternative stressed forms are encoded under the same lexical form with different ontolex:writtenRep, stress-labelled, properties;
 - There are cases when different pronunciations distinguish between two different words, in spite of their graphical identity: e.g., for lemma *afin_n*, the stress position differs for the meaning “Person in a legal affinity relationship” (*af'in*) and for the meaning “Short shrub, very branched, with oval-elongated leaves, light pink flowers and edible fruits, widespread in mountainous regions (*Vaccinium myrtillus*)” (*'afin*). For such situations, associations between senses/meanings and

¹³ <https://universaldependencies.org>

¹⁴ <https://universaldependencies.org/u/feat/>

pronunciation information would be important to encode. At this conversion step, we did not decide upon the best solution to encode this associations, but it is an essential issue to cover in future steps since it affects text-to-speech applications' ability to deal with such ambiguities. At the moment, all stressed versions are encoded consecutively under the same lexical form.

The LLOD RoLEX has 2,501,194 triples, with 361,974 `ontolex:Form` instances and 31,313 `ontolex:LexicalEntry` instances.

3 Publishing Romanian LLOD Resources

There are three ways to publish Linked Data resources [Verborgh, 2021], presented here with some of their features:

- *data dump*: an archive containing the Linked Data files is made available for download by those interested; an important advantage is that each user has flexibility in using the data; a disadvantage is that once downloaded, the resource is used in that version and any new versions released later must resume the processing flow from the user;
- *SPARQL endpoint*: the resource becomes available for querying, using the SPARQL language, by users who have access only to the information requested by the query phrases; the advantage is that they always have access to the latest version of the resource; SPARQL language requires costs (of resources);
- *Linked Data documents*: they allow access to per-topic data, always working with their latest version; by interconnecting resources, the query becomes accessible simultaneously in several resources.

Once converted to the Linked Data format, the Romanian resources were published (data dump) on a dedicated site¹⁵, with pages in Romanian and English. They can be freely downloaded and used by interested parties.

An Apache Jena Fuseki SPARQL server¹⁶ has been installed on the RELATE platform. It can upload RDF-Turtle files, similar to those produced in this project, and then allow them to be queried online by providing a SPARQL endpoint. The list of available resources is presented, the desired one can be selected and then the query can be launched.

The publication of L(O)D resources in accordance with the FAIR principles¹⁷ for scientific data management has led us to the LOD Cloud¹⁸ as a repository of

¹⁵ <https://www.racai.ro/p/llod/>

¹⁶ <https://relate.racai.ro/datasets/>

¹⁷ <https://www.go-fair.org/fair-principles/>

¹⁸ <https://lod-cloud.net/>

our resources: this is where metadata of valuable resources are recorded and where their interconnection with other resources is visible as arcs in the diagram. It is important for the development of language technologies and in languages with fewer resources to be as visible as possible, easy to find, while their interconnection with other resources increases their value by increasing the possibilities of use. Established in 2011 as a means “to measure and visualize the adoption of linked and open data within the linguistics community” [McCrae *et al.*, 2016], the Linguistic Linked Open Data Cloud¹⁹ is the result of an effort by the Open Linguistics Working Group²⁰ and contains metadata of resources that are, broadly speaking, relevant for the domain of linguistics. The metadata of the LLOD resources for Romanian we have converted are available in the LOD Cloud.

4 *Linking Resources*

One of the specific features of data in Linked Data format is their interconnectivity, which goes beyond their type, the language(s) described, even the domain to which they belong. This is a key feature of LD resources, as these links between entities from different resources allow cross-accessing information that is not present in a resource. At a cross-lingual level, RoWN is linked to PWN, while at a monolingual level, RoWN and RoLEX have been linked so far, with others to follow in the next phases of our project.

4.1 *RoWN – other Wordnets*

In the process of adapting RoWN to LLOD requirements [Barbu Mititelu *et al.*, 2020], the alignment of the synsets in RoWN to ILI²¹ was also made for 58,807 of the 59,348 synsets it contains. The alignments were automatically identified given the preexisting mapping of RoWN to PWN3.0. The link between a synset and the ID corresponding to the respective concept in the ILI list is established by the *ili* attribute, specifying in each synset a link to the corresponding concept through the *ili* property. Through the language-independent characteristic of ILI, RoWN is linked to any other wordnet that is, in its turn, linked to ILI.

4.2 *RoWN – RoLEX*

In Section 2.4, we mentioned sense information representation in the LLOD format of RoLEX and we will detail here what is meant by that. Sense information is not encoded in RoLEX per se, in the tabular format, but is added in the LLOD format as a means of interconnection with wordnet clusters like Global Wordnet

¹⁹ <https://linguistic-lod.org/llod-cloud>

²⁰ <http://linguistics.okfn.org/>

²¹ <https://github.com/globalwordnet/ili>

Grid or Open Multilingual Wordnet, and, implicitly, with RoWN (see above, Subsection 4.1).

The ILI/CILI index [Bond *et al.*, 2016], designed in the EuroWordNet project [Vossen, 1998] as a list of unique identifiers for all the synsets in all the participating wordnets, was used by us to link RoLEX to RoWN and to associate sense information to (some of) RoLEX entries. From this perspective, the LLOD version is an enriched resource if compared with the one in tabular format. Because RoLEX is just a list of words and it cannot provide context for the semantic disambiguation of its entries, each lexical entry is associated with a list of possible ILIs as retrieved from RoWN.

The procedure of linking RoLEX to RoWN through ILI is the following:

- for each lexical entry in RoLEX, uniquely identified through a lemma and a POS, recover from RoWN all synset IDs associated to synsets that contain the specific lemma and POS combination;
- for each recovered synset ID, identify the associated ILI in RoWN;
- extract a list of ILIs and add it to the corresponding lexical entry through the `wn:ili` property.

As we mentioned in Section 2.4, we can provide semantic disambiguation for words where pronunciation distinguishes between senses, and we intend to implement this disambiguation in a future step.

5 Conclusions

We have presented here Romanian resources that have been recently converted and published as LLOD ones, according to the current practices. The process is challenging at times, given the incomplete coverage the existing LLOD vocabularies have of the language descriptions. Two major challenges we have faced are connected with the representation of lexical stress in words and of the syllabification of word forms. Future work of our team will specifically focus on the enrichment of links between the Romanian LLOD resources, given that each contains the description of the language at a certain level, not an exhaustive one.

References

- Barbu, A.-M. (2008). Romanian Lexical Data Bases: Inflected and Syllabic Forms Dictionaries. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 1937–1941.
- Barbu Mititelu, V., Mitrofan, M. (2020). Romanian Resources in Linguistic Linked Open Data Format. In *Proceedings of the 15th International Conference “Linguistic Resources and Tools for Natural Language Processing”*, pages 7–16.

- Barbu Mititelu, V., Cristescu, M., Onofrei, M. (2019). The Romanian Corpus Annotated with Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 13–21, Florence, Italy. Association for Computational Linguistics.
- Barbu Mititelu, V., Irimia, E., Păiș, V., Avram, A.-M., Mitrofan, M., Curea, E. (2020). Romanian Resources in Linguistic Linked Open Data Format. In *Proceedings of the 15th International Conference "Linguistic Resources and Tools for Natural Language Processing"*, pages 29–40.
- Barbu Mititelu, V. (2018). Modern Syntactic Analysis of Romanian. In Ofelia Ichim, Luminița Botoșineanu, Daniela Butnaru, Marius-Radu Clim, and Veronica Olariu, editors, *Clasic și modern în cercetarea filologică românească actuală*, pages 67–78. Editura Universității "Alexandru Ioan Cuza".
- Bond, F., Vossen, P., McCrae, J.P., Fellbaum, C. (2016). CILI: the Collaborative Interlingual Index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57.
- Boroș, T., Pipa, S., Barbu Mititelu, V., Tufiș, D. (2017). A Data-Driven Approach to Verbal Multiword Expression Detection. PARSEME Shared Task System Description Paper. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pages 121–126, Valencia, Spain.
- Boroș, T., Dumitrescu, Ș.D., 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*, pages 171–179.
- Chiarcos, C., Făth, C. (2017). CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In J. Gracia, F. Bond, J. McCrae, P. Buitelaar, C. Chiarcos, and S. Hellmann, editors, *Proceedings of Language, Data, and Knowledge*, LNAI, pages 74–88, Springer.
- Ide, N., Pustejovsky, J. (2010). What does Interoperability Mean, anyway? Toward an Operational Definition of Interoperability for Language Technology. In *Proceedings of the Second International Conference on Global Interoperability for Language Resources*.
- Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*". Ph.D. thesis, Romanian Academy.
- Landthaler, J., Wärtl, B., Matthes, F. (2016). Unveiling References in Legal Texts-Implicit versus Explicit Network Structures. In *IRIS: Internationales Rechtsinformatik Symposium*, volume 8, pages 71–8.
- Leitner, E., Rehm, G., Moreno-Schneider, J. (2019). Fine-Grained Named Entity Recognition in Legal Documents. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham. Springer International Publishing.
- McCrae, J., Chiarcos, C., Bond, F., Cimiano, F., Declerck, T., de Melo, G., Gracia, J., Hellmann, S., Klimek, B., Moran, S., Osenova, P., Pareja-Lora, A., Pool, J. (2016). The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2435–2441.

- Mitrofan, M. (2019). *Extragere de cunoștințe din texte în limba română și date structurate cu aplicații în domeniul medical*. Ph.D. thesis, Romanian Academy.
- Păiș, V., Mitrofan, M. (2021). Towards a named entity recognition system in the romanian legal domain using a linked open data corpus. In *Workshop on Deep Learning and Neural Approaches for Linguistic Data*, pages 16–17, Skopje, North Macedonia, September.
- Păiș, V., Tufiș, D., Ion, R. (2020). A processing platform relating data and tools for Romanian language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 81–88, Marseille, France, May. European Language Resources Association.
- Păiș, V., Mitrofan, M., Gasan, C.L., Coneschi, V., Ianov, A. (2021). Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.
- Păiș, V. (2019). *Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language*. Ph.D. thesis, Romanian Academy.
- Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaite, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Escartín, C.P., Qasemi Zadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A., Walsh, A. (2018). Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Ramisch, C., Savary, A., Guillaume, B., Waszczuk, J., Candito, M., Vaidya, A., Barbu Mititelu, V., Bhatia, A., Iñurrieta, U., Giouli, V., Güngör, T., Jiang, M., Lichte, T., Liebeskind, C., Monti, J., Ramisch, R., Stymne, S., Walsh, A., Xu, H. (2020). Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 107–118, online. Association for Computational Linguistics.
- Sag I.A., Baldwin T., Bond F., Copestake A., Flickinger D. (2002) Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2002. Lecture Notes in Computer Science, vol 2276, pages 1–15. Springer, Berlin, Heidelberg.
- Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S.R., Eryigit, G., Giouli, V., van Gompel, M., HaCohen-Kerner, Y., Kovalevskaite, J., Krek, S., Liebeskind, C., Monti, J., Escartín, C.P., van der Plas, L., Qasemi Zadeh, B., Ramisch, C., Sangati, F., Stoyanova, I., Vincze, V. (2018). *PARSEME Multilingual Corpus of Verbal Multiword Expressions*, pages 87–147. Language Science Press, Berlin.
- Stan, A., Yamagishi, J., King, S., Aylett, M. (2011). The Romanian Speech Synthesis (RSS) Corpus: Building a high quality hmm-based speech synthesis system using a high sampling rate, *Speech Communication*, 53(3):442–450.
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J. (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the*

- Demonstrations at the 13th *Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April. Association for Computational Linguistics.
- Straka, M., Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.
- Straka, M., Hajic, 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)*, pages 4290–4297.
- Toma, Ș.-A., Stan, A., Pura, M.-L., Bârsan, 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. IEEE.
- Tufiş, D., Mitrofan, M., Păiş, V., Ion, R., Coman, A. (2020). Collection and annotation of the Romanian legal corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2773–2777, Marseille, France, May. European Language Resources Association.
- Tufiş, D., Barbu Mititelu, V. (2014). The Lexical Ontology for Romanian. In Nuria Gala, Reinhard Rapp, and Nuria Bel-Enguix, editors, *Language Production, Cognition, and the Lexicon*, volume 48 of *Text, Speech and Language Technology*, pages 491–504. Springer.
- Tufiş, D., Barbu Mititelu, V., Irimia, E., Păiş, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian. *Revue roumaine de linguistique*, LXIV(3):227–240.
- Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Niton, B., Ogrodniczuk, M., Pezik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiş, V., Tufiş, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J. (2020). The MARCELL legislative corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France, May. European Language Resources Association.
- Verborgh, R. (2021). Web fundamentals. Linked Data Publishing, <https://rubenverborgh.github.io/WebFundamentals/linked-data-publishing/>
- Vossen, P. (1998). Introduction to EuroWordNet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer.

Developing Linguistic Resources for Romanian Written and Spoken Language

OANA NICULESCU

Romanian Academy, Institute of Linguistics “Iorgu Iordan – Al. Rosetti”

oeniculescu@yahoo.com

Abstract

In this paper two research initiatives are presented in relation to the pending need for developing and exploring linguistic resources dedicated to the study of Romanian language. The first project deals with dialectal text retrieval from the Phonogramic Archive (AFLR) in terms of digitization and post-processing tasks. The second initiative is centred around the development and dissemination of an open access high quality speech corpus of contemporary standard Romanian. Both projects are carried out at the Romanian Academy Institute of Linguistics “Iorgu Iordan – Al. Rosetti” within the Department of Dialectology and Sociolinguistics.

1 Introduction

In this article we address the pending need for developing and exploring linguistic resources dedicated to the study of Romanian language, both written and spoken. The first part of the paper focuses on hand written dialectal document retrieval through a digitization project realized at the Bucharest Institute of Linguistics “Iorgu Iordan – Al. Rosetti” (Department of Dialectology and Sociolinguistics). When implementing this project granted by the Romanian Academy through the Digital Management Initiative for Safeguarding of the National Cultural Heritage, our main objective was to digitally recover and process the speaker files and filed notebooks belonging to the Romanian National Phonogram Archive (AFLR). Overall, 12.995 files were digitized. These documents are linguistically and culturally relevant from a transdisciplinary perspective, thus contributing to the advancement and dissemination of written resources for Romanian. In terms of spoken language resources, through our current postdoctoral research project, we aim to record, transcribe, align and explore a speech corpus of contemporary standard Romanian from a linguistic perspective. Both read speech and spontaneous speech are accounted for. This oral corpus will enable an in-depth examination of language variation, at the same time fostering new insights into speech therapy, language pathology and natural language processing. Various parts of the corpus can prove useful for designing perceptual experiments or developing materials for teaching Romanian as a foreign language. Through this open-source corpus, we can enable the investigation of a Romance language less

analysed on connected speech data, but with an increasing demand of digital learning tools.

The rest of the paper is organized as follows. Section 2 deals with Written Corpus Design in terms of background and motivation (2.1), corpus development (2.2.) through digital retrieval and digital processing activities, concluding with the utility and availability of the corpus (2.3). The Speech Corpus Design is presented in the third section, mirroring the structure of the previous section. The remaining part of the article is dedicated to conclusions and future research (section 4).

2 *Written Corpus Design*

2.1 *Background and Motivation*

Recording, transcribing and exploring the intricate aspects of the orality of peasant cultures remains at the heart of dialectal and ethnographic studies alike. Either stored on wax-cylinders, reel-to-reel tapes, cassettes or recent disk-storage devices and cloud-based services, the purpose of such oral collections is to preserve the collective memory, traditions, customs and folkloric productions of different communities. At the same time, these oral narratives uncover dialectal variety of the spoken language, functioning as a bridge between linguistics and folklore studies, especially in the context in which Romanian dialect literature is underrepresented [Vulpe, 1978].

One of the largest collections of dialectal texts can be found at the Romanian Academy Institute of Linguistics “Iorgu Iordan – Al. Rosetti” [Marin, 1996]. Established in the late 1950s at the initiative of Al. Rosetti after the initial phonograph recordings were lost during World War II, The Phonogramic Archive of the Romanian Language (also referred to as AFLR) gathers authentic and representative samples of rural oral productions pertaining to all Daco-Romanian subdialects, both in the country as well as across the borders. The Archive is a vast, multifaceted repository of the community’s collective memory and identity. The interviews are well-designed for eliciting spontaneous speech, thus drawing closer to the source material under examination.

During the gathering of the dialectal texts, researchers saved the data on two formats. First, the audio recordings were stored on reel-to-reel tapes which were later used for transcribing the interview. Second, the speaker meta-data and the list of conversation topics, dialectal terminology and pronunciation variants written in narrow phonetic transcriptions were gathered in what is referred to as “village notebooks”. AFLR holds over 1800 tapes and 950 notebooks. The digitization of the collection lies with the members of the Dialectology and Sociolinguistics Department from the Bucharest Linguistic Institute. Through one of the fundamental preservation programs at the Institute, developed as an in-house protocol, the focus is on digitally retrieving and processing the audio files

so as to facilitate the further transcription of dialectal texts. In this context, we initiated a complementary digitization project (between 2019 – 2021) built around the recovery of the hand written texts present in the village notebooks.

The motivation of this initiative is twofold. On the one hand, the digitization of the Archive in its entirety remains one of the key assignments for the Department handling the Archive, in particular, and for the Linguistic Institute, in general. As such, this digitization initiative comes as a relevant tool for ensuring the preservation of the Archive. On the other hand, all documents from the collection, either audio or text, risk being lost due to internal and external risks. As a way to ensure the preservation of the texts, the digitization project was launched, being the first of its kind carried out at the Department. The challenges of developing and implementing this project concern setting up a methodology and roadmap for compiling, digitizing and processing the corpus. The ways in which we addressed these issues are presented in the following (sub)sections.

2.2 Corpus Development

The development of the written corpus is done in 3 stages implemented through 8 activities as described below.

STAGE 1 CORPUS COMPILATION

The main objective at this stage is centred around file identification, indexation and proper storage of the hand written documents from the Archive.

Activity 1: Archive organization. The objective set for this activity deals with establishing and organizing the intended material for the digital retrieval process (Task 1.1 Document search; Task 1.2 Document arrangement).

Activity 2: File classification. The goal of this activity is to classify the material identified in the previous stage (Task 2.1 Setting up an indexing system; Task 2.2 Document numbering; Task 2.3 Document indexation).

Activity 3: File storage. The objective of this step is to ensure a proper storage for the textual documents (Task 3.1 Optimizing the storage space; Task 3.2 Document physical storage).

STAGE 2 CORPUS DIGITIZATION

The main goal at this stage is to digitally retrieve and safely storage as a digital document the files from the Archive.

Activity 4: Developing the procedure and training. The purpose of this activity is to develop and implement, through personnel training, a coherent and time effective procedure for document digitization (Task 4.1 Document preparation; Task 4.2 Task effective digital retrieving procedure).

Activity 5: Digitization. The aim of this activity is to digitize the hand written documents from the Archive based on the procedures developed within the

previous activity (Task 5.1 Document scanning; Task 5.2 Document uploading and storage on local external drive).

STAGE 3 CORPUS PROCESSING

The objective at this stage is to process, in terms of image editing, conversion and repository, the digitized files from the previous stage.

Activity 6: Image editing. The purpose of this activity is to digitally edit the scanned files (Task 6.1 Image adjusting (alignment, cropping); Task 6.2 Image retouching (background noise removal, spot healing, color and lighting correction)).

Activity 7: Image conversion (from .jpg/ .png to high quality .pdf). The purpose of this activity is to have the entire digitized notebook in one single file that will be used to create the flipbook. Another motivation of the conversion is to have all of the image files stacked together into a single file to avoid cluttering of the digital storage space (Task 7.1 From image to portable document file; Task 7.2 From portable document file to flipbook).

Activity 8: Document processing. The scope of this current activity is to set up a preliminary database of the digitized documents (Task 8.1 Setting up a system for easy digital file tracking; Task 8.2 Deciding on the fields to incorporate in the database; Task 8.3 Retrieving the relevant information from the scanned documents; Task 8.4 File backup on external hard drives and cloud storage).

2.3 Digital Retrieval

This digitization initiative is in line with the Romanian Academy Digital Management project for the national cultural heritage and has benefited from the Academy Grants for Humanities. Although the project was initially developed so as to recover the informant files (which were entirely organized, scanned and processed), we later extended the digitization program to include the phonogramic files (Figure 1 and Figure 2). The reason behind this decision lies, on the one hand, with the invaluable (extra)linguistic data found on these documents, relevant especially for dialectal and lexical analyses, and, on the other hand, with preventing the further degradation or possible loss of the collection due to internal and external factors.

The cardboard speaker's files have a size of 21 x 15 cm (8.27 x 5.91 inches), while the phonogramic records are of 15 x 21 cm. Due to the frail nature of the documents, each file was digitally recovered using the single sheet scanning method compatible with the flat-bed scanner. A Canon Lide 400 Scanner was used (scanner type – flatbed, sensor type – CIS, general format scanner A4, scan mode – colour, scan resolution (DPI) – 4800 x 4800). The resulting scanned images were edited using a professional photo editing software.

At the end of the digital retrieval process, 12.995 files were recovered. The speaker files comprise 4943 documents, and are distributed as following,

according to the county where the interviews took place: AB (230), AR (10), AG (17), BH (39), BN (49), BT (23), BR (71), BV (322), BZ (102), CL (68), CS (174), CJ (102), CT (126), CV (46), DB (162), DJ (8), GL (112), GR (41), GJ (56), HR (60), HD (n 75), IL (263), IS (33), IF (105), MM (45), MH (25), MS (320), NT (44), OT (107), PH (148), SJ (42), SB (298), SV (49), TR (245), TM (85), TL (113), VL (128), VS (19), VN (64). An additional 760 files corresponding to sociolinguistic interviews for AFLR carried out in schools were also digitized. The remaining 8052 documents, recovered in addition to the initial research proposal, constitute 324 of the village notebooks (i.e., 30% of the entire collection related to fieldwork notes). The next step in our project is related to processing the digitized documents as will be described in the following section.

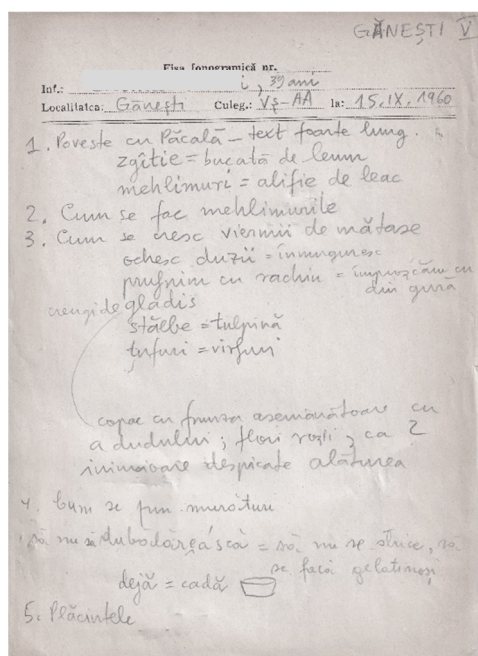


Figure 1. Phonogramic file retrieved from the village notebook No. 357 (Galați, Moldavia, September 1960)

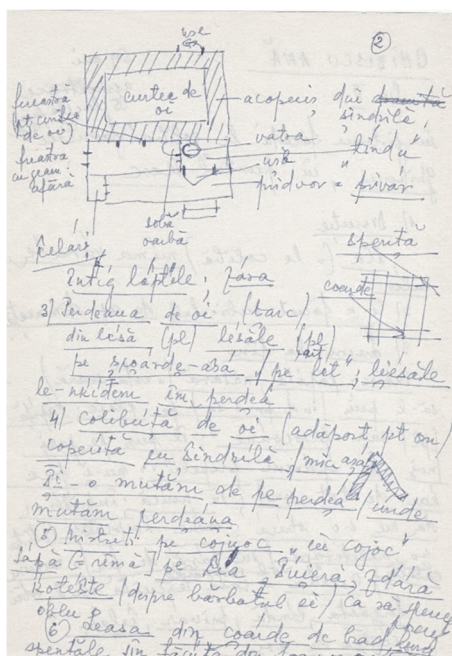


Figure 2. Phonogramic file retrieved from the village notebook No. 690 (Alba, Transylvania, August 1989)

2.4 Digital Processing

All digitized files went through image editing and conversion. Alongside the digital storage of the archival documents, another key aspect, which is fundamental from our perspective, is to have quick and easy access to the relevant data present on the files. Thus, the final step of our digitization program was to set up a local database. After going through the entire set of documents, we

decided on using a structure made up of 20 fields of interest grouped in 4 sections further described below.

Section 1 – Document Type

- Speaker files: metadata regarding the place and date of the recordings, initials of the interviewers, the surname and (nick)name of the speaker, age and education, occupation (primary and secondary), possible travels outside the village (either work related or military service), details regarding family status (husband, wife, children). and observations (mainly related to recording conditions).
- +/- Photograph: whether the document has attached a photograph of the speaker (usually in front of the house, dressed in traditional clothing).
- Phonogramic file: document containing linguistic data related to conversation topics, pronunciation variants accompanied by narrow phonetic transcriptions, glossaries and, in rare cases, drawings made by hand in order to illustrate various dialectal words (for more information on this topic, see [Niculescu and Manta, 2020]).
- Observations: usually, the speaker file is correlated with the phonogramic file; in case this correlation is absent (meaning that one of the files is missing from the Archive), the resulting mismatch is highlighted.

Section 2 – Place and Date of the AFLR Recordings

- Village: the (updated) name of the village where the recordings took place; it is marked if the village is also the commune seat.
- Commune: the (updated) name of the commune where the interview took place.
- County: the (updated) county where the interview took place.
- Year: year of the recording.
- Month: month of the recording.

Section 3: Speaker Information

- First Name: name given upon birth, in the files appears sometimes together with the middle name
- Second Name: family name; for the female speakers it is not marked whether it is their maiden name or their husband's surname after marriage.
- Nickname: more common in male speakers, sometimes appears in phonetic transcription alongside the surname.
- Age: the age of the speaker at the moment when the recording was made.

- Sex: the speaker's gender.
- Symbol: based on an inventory system developed by the AFLR researchers [Șuteu, 1958] according to which speakers were attributed figures from I to VIII (male speakers were given even numbers, while female speakers received odd numbers), I and II correspond to the first generation of speakers, namely children between 9 and 12 years old, III and IV represent the second generation, adolescents between 18 and 22 years old, V and VI characterize the third generation of speakers, namely adults within 35 and 45, while VII and VIII showcase the fourth generation of speakers, that is adults over 60 years old.
- Education: literate (number of years spent in school) or illiterate.
- Occupation: the primary and secondary occupation of the speaker at the moment when the AFLR interview was made; for the male speakers (IV, VI, VIII) the most common professions recorded are shepherds, farmers, beekeeper, cattlemen, carpenters, builders and bricklayers, railroad worker, factory worker, day labourer, woodsman, hunter, while for female subjects (III, V, VII) the primary occupation is marked as housewives, followed by collectivists and cleaning lady.

Section 4 – Correlating Text with Audio

- Reel-to-reel tape: number and length; each track is between 30 to 90 minutes long; correlating the number of the reel-to-reel tape with the number of the village notebook in which the informant and the phonogramic files are stored. For instance, the texts from the village notebooks 357 and 690 (Figure 1 and Figure 2) are stored on the tapes number 655 and 1598, respectively.
- State of digitization: whether or not the tape has been digitized.
- Village network: we mark the cartographic point number from the digitized files based on “Atlasul lingvistic roman pe regiuni. Sinteză” (ALLR) [The Romanian Regional Linguistic Atlas. Synthesis]: Banat (from cartographic point number 1 to point number 100), Crișana (101 – 220), Maramureș (221 – 240), Transilvania (241 – 460), Moldova and Bucovina (461 – 670), Muntenia and Dobrogea (671 – 900), Oltenia (901 – 998), Basarabia, Northern Bucovina and Transnistria (01 – 0224). It is important to note that although the AFLR closely followed the same network on which the national dialectal atlases are based on, there various recordings pertaining to regions outside of this network, situation marked in the database with an asterisk.

2.5 *Availability and Utility of the Written Corpus*

At the present moment, the digitized files remain an in-house resource of the Linguistic Institute. The digitization project was carried out at the Department of Dialectology and Sociolinguistics between 2019 – 2021. Even though the financial support received through the Romanian Academy Grants for Humanities has come to an end in October 2021, this digitization initiative is set to be resumed in the following years, leading to the entire recovery of the textual documents within AFLR. The plan is to continue scanning and processing the remaining notebooks and add the data to the existing database. The scanned documents will also be transcribed in an effort to make the data compatible with Handwriting Recognition systems. After the digitization process is finalized, we can proceed to making the data available through a Creative Commons license in agreement with the European norms. Another option is to distribute the material through the SOLIROM platform dedicated to “Instruments and digital academic linguistic resources for the Romanian language” (<https://solirom.ro/>). This step remains to be further analysed. For more information and updates on this subject, please visit the official website of the Linguistic Institute “Iorgu Iordan – Al. Rosetti” (<https://lingv.ro/>). Various linguistic resources were also developed throughout this project, such as video tutorials, handouts, audio recordings, among other things, and are available on the aforementioned website (on this topic, see [Niculescu, 2021]).

Alongside the intrinsic dialectal purpose and use of these archival documents, the material is relevant from a trans- and interdisciplinary perspective. The speaker files contain significant data for social-onomastic studies, especially the field dedicated to nicknames. Also, these rich data stored on the files are valuable for historians and sociologists alike. Ethnographic studies can be carried out based on the black and white photographs attached to documents. This perspective is likewise relevant in the context of broadening The Romanian Peasant Museum photograph collection. The phonogramic files have a direct impact on extending the lexical and lexicographic resources dedicated to Romanian. The results of this digitization project can also have an impact on developing and extending the tools for natural language processing especially when working on dialectal data and retrieving hand written documents.

3 *Speech Corpus Design*

3.1 *Background and Motivation*

In the present context of inter- and trans- disciplinary research, when Digital Humanities are bringing forth new tools and methods into the study of language variation, annotated corpora prove to be of high relevance in linguistics in general, and more so in phonetic investigations [Vasilescu *et al.*, 2014]. Modern phonetic research has drawn close to speech technologies. As a result, the key aspect of

working at the interface between experimental phonetics and language technologies is that it can help linguists investigate the relation between synchronic variation and sound change, leading to a better understanding of the evolution within a language [Ohala, 1996].

In 2012, when the Romanian language was presented in the Digital Age [Trandabăţ *et al.*, 2012], it appeared to be an underrepresented language, with a rather limited representation in fields such as large corpora analysis. At that time, national speech annotated corpora for Romanian were scarce and rather difficult to access [Mîrzea Vasile and Timotin, 2015], needing permission from the coordinator. There were certain drawbacks which proved harder to overcome such as poor audio quality, vague metadata, unstructured interview making the data challenging to compare, ambiguous policy of data collection and speaker consent. A well-documented and detailed account regarding Romanian corpora from a linguistic perspective can be found in Mîrzea Vasile [2017]. Since then, numerous initiatives for developing resources and tools for natural language processing were carried out at RACAI (Research Institute for Artificial Intelligence “Mihai Drăgănescu”), alongside the Natural Language Processing Group from the Faculty of Computer Science of the “Alexandru Ioan Cuza” University. The Reference Corpus for Contemporary Romanian Language (CoRoLa) [Mititelu *et al.*, 2018] was launched. The Speed research laboratory likewise plays an important role in building and improving automatic speech recognition systems for Romanian [Cucu, 2011; Cucu *et al.*, 2014], also developing the Read Speech Corpus [Georgescu *et al.*, 2020]. The natural language platform RELATE [Păiş *et al.*, 2019; 2020], which evolved from the TEPROLIN platform [Ion, 2018], designed at the Institute for Artificial Intelligence of the Romanian Academy, is another key resource facilitating the access to various tools for the Romanian language. The bimodal corpus for Romanian created through the CoBiLiRo project [Cristea *et al.*, 2019] marked another important resource. The Romanian Speech Synthesis corpus [Stan *et al.*, 2011] alongside the SWARA Speech Corpus [Stan *et al.*, 2017] and the Mara corpus [Stan *et al.*, 2021], licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, are other noteworthy resources. It is important to mention that a recent Learner Corpus for Romanian is also under development (see the LECOR project presented in [Mîrzea Vasile, 2020]).

In this context, the ROC-lingv project comes as complementary initiative, with the objective of providing an open access high quality oral corpus of contemporary standard Romanian designed to follow the EU recommendations, in agreement with the GDPR norm. The data acquired are manually and automatically segmented. By making the corpus available to the general audience, one of the main scientific contributions of the ROC-lingv project is to provide modern resources for the study of Romanian, thus allowing linguistic comparisons with other Romance languages and opening new paths of research. Our postdoctoral

research project, financed by UEFISCDI (2020 – 2022), is carried out at the Linguistic Institute in Bucharest and it is mentored by Prof. Andrei A. Avram from the University of Bucharest. The challenges we face when conducting this project range from experiment design, optimal linguistic coverage and adequate transcript system, to data processing and alignment. The ways in which we address these issues are presented in the following (sub)sections.

3.2 *Corpus Design*

The project is designed in 2 stages. The first stage, *Corpus design*, is defined by data gathering and processing, while the second stage, *Studying linguistic variation*, is centred around data exploration and dissemination. In what follows, each activity will be described.

STAGE 1 CORPUS DESIGN

The ROC-lingv project aims to deliver an open access corpus of spoken contemporary standard Romanian with the key purpose of stimulating innovative research at the interface between phonetics and phonology within the national research program. This objective will be accomplished through corpus acquisition, transcription, annotation and processing tasks balanced across a feasible and coherent work schedule. Our key intermediate goals are assuring a balanced linguistic coverage, building an efficient transcribing and annotating system, conveying an optimal corpus design, and correlating the manual segmentations with the forced alignment.

Activity 1: Data collection. The objective set for this activity deals with corpus acquisition (Task 1.1 Experimental setup; Task 1.2 Recordings).

Activity 2: Data processing. The purpose of this activity is to go through the data and render the recordings available for linguistic studies (Task 2.1 Audio and transcription processing; Task 2.2 Segmentation processing; Task 2.3 Output post-processing).

STAGE 2 STUDYING LINGUISTIC VARIATION

The second part of the project focuses on exploring the corpus from a linguistic perspective and outlining a taxonomy of connected speech phenomena in contemporary standard Romanian. The aim of the ROC-lingv project consists in developing new methods and tools for acoustic analysis and integrating them into a coherent and well-designed framework. This way we can link acoustic findings to large data analysis, thus improving national phonetics research.

Activity 3: Statistical overview. The goal of this activity is to provide a unique and up-to-date statistical computing of the data acquired in the previous stage of the project. This activity is be carried out through two tasks (Task 3.1 Phonotactics; Task 3.2 Word-frequency lists).

Activity 4: Sound change and phonetic variation. A forced aligned oral corpus is valuable since it can capture the inherent dynamic nature of language in use, aiding to the understanding of phonetic variation and sound change. As such, the objective of this activity is to describe the data from a linguistic perspective (Task 4.1 Preliminary typology of relevant phonetic variation in Romanian; Task 4.2 Testing various linguistic hypothesis).

3.3 *Data Collection*

The ROC-lingv project has two main objectives.

- Recording, transcribing and aligning an oral speech corpus of contemporary standard Romanian for linguistic purpose. Both controlled (read speech) and spontaneous speech (monologues) are accounted for.
- Proposing a typology of relevant phonetic variation in Romanian. We will be looking at inter- and intra- speaker variability with respect to various reduction processes found in connected speech. This analysis opens up various discussions with respect to modelling gradient phonetic and phonological phenomena, leading to an in-depth analysis of linguistic variation and sound change

In terms of speaker selection, we limit our analysis to 12 subjects (6 males, 6 females), adults (30 to 40 years of age), representative of the southern dialect, on which the standard language is based on. We are not documenting different accents across the country. In this context, due to resource constraints, both in terms of time (2 years) and personnel (limited the project director), we must also acknowledge the limitations of our proposal, both in terms of time-management and personnel resources. The ROC-lingv project is meant to be a stepping point in modern acoustic research, showcasing the advantages of working at the interface between acoustics and speech technologies.

All recordings are laid out in two sessions, corresponding to the elicitation of controlled speech (experiment 1 – word list) and spontaneous speech (experiment 2 – monologue). The corpus is recorded in the Phonetic Laboratory at the Bucharest Linguistic Institute. The recording booth within the laboratory is equipped with a studio quality microphone for increased audio fidelity, connected to a PC through an external audio interface. All recordings are carried out within Praat (<http://www.fon.hum.uva.nl/praat/>), an open-source acoustic program developed by Paul Boersma and David Weenink, with the standard settings (44100 Hz, Channels – Mono), but extending the buffer size (Praat > Preferences > Sound recording preferences > Buffer size (MB)). No compression is added to the audio file. In the present context, it is important to mention the fact that speakers are not wearing surgical masks or respirators (FFP2 or N95 masks). We provide speaker-metadata for each recording and all participants willingly sign an agreement following the GDPR norm.

During the first experiment, the participants receive a word-list with the target word placed in a carrier sentence (“Zic ____ tare”; “I say ____ loud”). Participants read the randomized stimuli at a normal speech rate, going through the entire set five times. The main acoustic analyses derived from this experiment are related to VOT (voice onset time) measurements for stop consonants, fricatives and affricates COG measurements and duration of the friction noise, as well as an in-depth study on Romanian vowels. Experiment 2 is aimed at eliciting spontaneous speech through a monologue task (does not involve speaker overlap, allows for full data recovery, being also faster to transcribe and annotate). The monologues share the same topic discussions, making the data comparable. Based on the processed data up to this moment, the monologues have a mean duration of 45 minutes.

At the beginning of the interview, each candidate received a printed handout (Figure 3) containing the three main conversational topics: (1) Pursuits (refers to present activities) – in my spare time, travelling, what I like, what I dislike; (2) Memories (refers to past activities) – from childhood, sophomore years, life lessons; (3) Forthcoming projects (refers to future activities) – personal projects, professional projects. Fifteen minutes are given prior to the recording session during which participants can write down their ideas in relation to the topics proposed using only key-words and short phrases. The experiment is not intended to elicit read speech. During the interview subjects can move freely from one topic to another, no order is pre-imposed, the goal being to emulate natural, (semi)spontaneous speech and, at the same time, allow for future cross-examination based on comparable data collection.

3.4 Data Processing

In terms of data processing, the goal is to prepare the data recorded previously for linguistic analysis. The tasks needed to achieve this objective range from data labelling, matching the orthographic transcription with the broad phonological rendering of the recordings, designing the phonetization rules and pronunciation dictionary, to post-processing the outputs generated from the alignment. The recordings went through audio processing, name anonymization and individual file storage. As for the manual transcriptions, due to time restraints, we will not be delivering a verbatim. As an alternative, we propose to pair the orthographical transcription with a broad phonological transcript. Transcription conventions will be available on the project website. In the future, as a natural continuation of the project, syntactic and prosodic annotations will be added (the latter adapted to ToBI (Tones and Break Indices – initially developed for American English, see [Silverman *et al.*, 1992]) by [Jitcă *et al.*, 2012]).

All recordings are labelled and segmented in Praat. For each .wav file, a corresponding TextGrid is generated with the orthographical transcription on the first tier, followed by the phonological annotation on the second tier. In the case

of the second experiment, we focus on manual transcriptions paired with automatic forced alignment of spontaneous data. The automatic speech recognition system will take as input the speech files and their corresponding manual transcriptions. The words from the transcription are decomposed into phonemes using the pronunciation dictionary and the associated pronunciation variants defined in the lexicon. Each phoneme is then forced aligned with an audio segment. The corpus will be aligned with a system described in [Vasilescu *et al.*, 2014; Vasilescu *et al.*, 2015]. The files generated from the forced alignment are post-processed using Excel Kutools and R, a data analysis environment used for statistical computing and data visualization (<https://www.r-project.org/about.html>). Data cleaning and name anonymization are conducted through Adobe Audition CC.

In terms of exploring the corpus, one of our research interests deals with providing an updated phonotactic analysis. Romanian benefits from only one in depth phonotactic account dating back to more than half a century ago [Roceric-Alexandrescu, 1968]. In this context, the ROC-lingv project sets to narrow the linguistic gap and deliver an updated phonotactic description by correlating the statistical results based on written corpora [Roceric-Alexandrescu, 1968] with the outputs generated from our connected speech data. Another task regarding the statistical overview of the recorded data deals with word-frequency lists organized according to lexical coverage, part of speech, number of characters. These results can prove useful especially for teaching Romanian as a second language.

Once the post-processing stage of the data alignment has been successfully completed, the ROC-lingv project can advance by compiling a taxonomy of relevant phonetic variation in Romanian. In other words, we will be looking at inter- and intra- speaker variation in terms of contextual coarticulation effects such as consonant cluster reduction, secondary palatalization, among others. The ROC-lingv project addresses a wide variety of scientific issues relevant for both linguistic and ASR applied research. Based on the oral corpus, numerous studies of variation in continuous speech can be carried out, enabling scholars from varied backgrounds and with different research questions to test various linguistic hypothesis. Our contribution to the field would be innovating current research techniques within phonetic analyses and offer a new insight on a European language less studied on connected speech data, but with an increasing demand of digital learning tools.

3.5 Availability and Utility of the Speech Corpus

Users are informed with respect to the speech corpus on the project website (https://lingv.ro/pn-iii-p1-1-1-pd-2019-1029_roc-lingv/) hosted by the Bucharest Linguistic Institute. The scientific reports are also available alongside the dissemination of the results (through conferences and articles). Information regarding labelling conventions, the IPA for standard contemporary Romanian as

well as the key steps to working with TetxGrid files in Praat will be uploaded on the website. Both pdf files with video tutorials will be made available to the broader audience regarding the ways in which the corpus can be explored and used for linguistic analyses. All these steps are taken in order to make the experience of downloading and working on the open-access corpus more user friendly

Since there is a high demand for spontaneous speech data (both from natural language processing as well as linguistic standpoint), the main advantage is that the speech corpus is not developed so as to address a specific research interest, rather it was designed in order to cover variation in language use. The ROC-lingv project is addressed to a wide variety of users. From linguists and language learners to speech therapists or technicians, this project answers numerous research questions. From a linguistic perspective, the oral corpus is equally relevant and allows for an in-depth examination of language variation ranging from prosody, semantics and morphology, to syntax and pragmatics. The data acquired contributes to general comparative Romance linguistics, a domain where the Romanian language is often misrepresented. The word frequency lists and the lexical coverage derived from the elicitation of spontaneous speech is suited also for language teachers and students. It can help develop materials for teaching Romanian as a foreign language and generate modern tasks in assessing Languages for Special purposes, an area still under development [Biriş, 2012]. The results from the controlled experiment can foster new insights into speech therapy and language pathology. Finally, our project can prove useful in developing linguistic tools designed specifically for Romanian data, without having to model pronunciation variants from other Romance languages.

4 *Conclusions and Future Work*

The article presents two research projects for developing linguistic resources for Romanian written and spoken language. In terms of the written corpus, the initiative will be furthered developed by digitally retrieving all textual documents from AFLR, extending the database and making the data available for researchers. There is also an interest in engaging in collaborative projects with various Linguistic Institutes in the country. As we could observe during the implementation of the Romanian Academy Grants for Humanities, there is a general interest for digitizing archival documents and thus contributing to various resources and tools for the Romanian language. Future work in relation to speech corpus development and dissemination is related to adding new levels of annotation, allowing for POS marking, extending the recordings to speakers from different age groups and regions.

Acknowledgements

This work is supported by a grant of the Romanian Ministry of Education and Research, CNCS-UEFISCDI, project number PN-III-P1-1.1-PD-2019-1029, within PNCDI III.

References

- Barbu Mititelu, V., Tufiş, D., Irimia, E. (2018). The Reference Corpus of Contemporary Romanian Language (CoRoLa). In Proceedings of the 11th Language Resources and Evaluation Conference – LREC’18, Miyazaki, Japan, European Language Resources Association (ELRA).
- Biriş, G. (2012). Vocabularul minimal al limbii române: stadiul cunoştinţelor, perspective. In *Noi perspective în abordarea românei ca limbă străină / ca limbă nematernă*, Elena Platon, Antonela Arieşan (ed.), Cluj-Napoca, Casa Cărţii de Ştiinţă, pages 173–18.
- Cristea, D., Pădurariu, C., Boghiu, Ş., Gifu, D., Onofrei, M., Trandabăţ, D., Pistol, I.C., Bibiri, A.D. Scutelniciu, A. (2019). The COBILIRO project: building and distributing a bimodal corpus for Romanian language. In Proceedings of the 14th International Conference “Linguistic Resources and Tools for Natural Language Processing” 2019, pages 13–24, Cluj-Napoca, Romania.
- Cucu, H. (2011). Towards a speaker-independent, large-vocabulary continuous speech recognition system for Romanian, PhD Thesis, University “Politehnica” of Bucharest.
- Cucu, H., Buzo, A., Petrică, L., Burileanu, D., Burileanu, C. (2014). Recent improvements of the Speed Romanian LVCSR system. In 10th International Conference on Communications (COMM). IEEE 2014, pages 1–4.
- Ion, R. (2018). TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018), November 22-23, 2018, Iaşi, Romania.
- Jitcă, D., Apopei, V. Păduraru, O. (2012). The RO-TOBI annotation system and the functional analysis perspective of the Romanian intonation. Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of The Romanian Language" 8-9 December 2011, pages 3 – 10, ISSN 1843-911X.
- Georgescu, A.-L., Cucu, H., Burileanu, C. (2020). RSC: A Romanian read speech corpus for automatic speech recognition. In Proceedings of the 12th Language Resources and Evaluation Conference 2020, pages 6606–6612, Marseille, France.
- Marin, M. (1996). Arhiva fonogramică a limbii române (După 40 de ani). *Revista de lingvistică şi ştiinţă literară*, 1: 41–46.
- Mîrzea Vasile, C., Timotin, E. (2015). Le roumain contemporain. In Iliescu, Maria, Eugen Roegiest (eds), *Manuel des anthologies, corpus et textes romans*, Berlin, De Gruyter Mouton (Manuals of Romance Linguistics 7), pages 612–632; ISBN: 978-3-11-033313-8.
- Mîrzea Vasile, C. 2017. Corpusurile de limba română şi importanţa lor în realizarea de materiale didactice pentru limba română ca limbă străină. *Romanian Studies Today*, 2017(1): 74–95.

- Mîrzea Vasile, C. (2020). The LECOR Project. A Presentation. In Proceedings of the 15 International Conference "Language Resources and Tools for Natural Language Processing, online, 14-15 December 2020, Iași: Editura Universității A. I. Cuza (2020): 37–46.
- Niculescu, O., Manta, M. (2020). O altă perspectivă asupra anchetei dialectale: situația schițelor și a desenelor din însemnările de teren pentru AFLR. Oral communication delivered at Cel de al 19-lea Simpozion internațional de dialectologie, 10-11 September, Institutul de Lingvistică și Istorie Literară „Sextil Pușcariu”, Cluj-Napoca.
- Niculescu, O. (2021). Resurse electronice dezvoltate în cadrul proiectului MIDVAL. Oral communication delivered at Atelierul on-line de lingvistică – București, 2nd edition, 29 September – 1st October 2021.
- Ohala, J. (1996). The relation between sound change and connected speech processes. *Institute für Phonetik und digitale Sprachverarbeitung*, pages 201–206.
- Păiș, V., Tufiș, D., Ion, R. (2019). Integration of Romanian NLP Tools into the RELATE Platform. In Proceedings of the 14th International Conference “Linguistic Resources and Tools for Natural Language Processing”, pages 181–192, Cluj-Napoca, Romania.
- Păiș, V., Ion, R., Tufiș, D. (2020). A Processing Platform Relating Data and Tools for Romanian Language. In Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020), pages 81–88.
- Roceric-Alexandrescu, A. (1968). *Fonostatistica limbii române*, București: Editura Academiei Române.
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M. (1992). ToBI: A standard for labeling English prosody. In *ICSLP*, pages 867–870.
- Stan, A., Yamagishi, J., King, S., Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication*, 53(3): 442–450.
- Stan, A., Dinescu, F., Țiple, C., Meza, Ș., Orza, B., Chirilă, M., Giurgiu, M. (2017). The SWARA speech corpus: A large parallel Romanian read speech dataset. In 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). IEEE, pages 1–6.
- Stan, A., Lőrincz, B., Nuțu, M., 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), IEEE, pages 85–90.
- Șuteu, V. (1958). Arhiva fonogramică a limbii române. *Fonetică și dialectologie*, (1): 211–219.
- Trandabăț, D., Irimia, E., Barbu Mititelu, V., Cristea, D., Tufiș, D. (2012). *The Romanian Language in the Digital Age*, META-NET White Paper Studies, Springer.
- Vasilescu, I., Vieru, B., Lamel, L. (2014). Exploring pronunciation variants for Romanian speech-to-text transcription. *Proceedings of SLTU-2014*, pages 161–168.
- Vasilescu, I., Dutrey, C., Lamel, L. (2015). Large scale data based investigations using speech technologies: the case of Romanian. *Proceedings of the 8th Conference on Speech Technology and Human-Computer Dialogue (SpeD 2015)*, Bucharest, October 14-17, pages 1–6.
- Vulpe, M. (1978). Romanian Dialectology and Sociolinguistics. *Cahiers de linguistique théorique et appliquée*, 23 (1-4): 293–328.

Digital Resources Developed Within the BIBLIO-MLRom Project

COSMINA-MARIA BERINDEI

“Sextil Pușcariu” Institute of Linguistics and Literary History
Romanian Academy – Cluj-Napoca Branch
cosmina.berindei@academia-cj.ro

Abstract

The project Digital Library of the Romanian Language Museum's Questionnaires' Corpus. Integrated System of Metadata Management (BIBLIO – MLRom) aimed to conceive a tool able to manage the answers to Romanian Language Museum's indirect inquiry, consisting of eight questionnaires sent between 1922 and 1938, after the Great Union, in the Romanian villages.

The assumed purpose of the Romanian Language Museum's linguistic inquiry was to collect data on spoken Romanian in order to stock them for using both as a lexicographical resource for the Romanian Language Dictionary and as an experiment for the Romanian Linguistic Atlas which was a very important project in Sextil Pușcariu's view.

The inquiry was done using eight thematic questionnaires: Questionnaire I. The horse [1922]; Questionnaire II. The house [1926]; Questionnaire III. The thread [1929]; Questionnaire IV. Names for places and people [1930]; Questionnaire V. Sheepfold, shepherding and milk processing [1931]; Questionnaire VI. Beekeeping [1933]; Questionnaire VII. Musical instruments [1935]; Questionnaire VIII. Food and drinks [1937]. The questionnaires were spread with the help of church administration units and school inspectors. The eight questionnaires totalized 1598 questions and there were received, according to the register of documents, 1719 answers.

The tool developed in the project provides quick access to documents in .pdf format, using the Calibre platform, an e-book library manager, which has been adapted for this type of document. Thus, upon entry into the digital library, each document was provided with descriptive metadata. Based on the metadata, this tool offers the possibility to carry out quantitative research on the fund, on the geographical distribution of the answers received to the inquiry, the occupation of correspondents or other issues.

Our paper presents the digital library and the way in which it can be used to carry out quantitative research on the fund.

1 Introduction

The digitization of the language archives generated in the second half of the nineteenth and early twentieth centuries represents a challenge at present, for the specialists in the field of both linguistics and IT. The general goal of the digitization process is to introduce the archives into the broad circuit of research and more specifically to make them lexicographic sources accessible for retrieval for the *Romanian Language Dictionary* which is conceived as a thesaurus. Thus, this paper draws attention to an extremely appealing archive collection, containing the answers to the *Romanian Language Museum's Questionnaires* established between 1922 and 1939. The creation of this collection within the Romanian Language Museum was conceived as a *lieu de mémoire*, a place where the Romanian collective memory was to be found [Berindei, 2019: 71–74]. Its founder, Sextil Pușcariu, who also helped found the Romanian university in the capital of Transylvania, aimed to make the dialects of the Romanian language part of the national heritage. The Romanian Language Museum that Sextil Pușcariu founded was the first research institute in the university in addition to the Faculty of Letters and Philosophy and was founded by the decision of the Governing Council of August 27, 1919 as "a place of original research" and a means "to make science available to the masses" [Pușcariu, 1921: 1–2].

At the Romanian Language Museum, Sextil Pușcariu wanted to accomplish a piece of work that was fundamental to the Romanian culture, the *Romanian Language Dictionary*, a project entrusted to him by the Romanian Academy in 1906 and which has continued to be written in Cluj-Napoca [Haja *et al.*, 2005: 17]. Another piece of work that he wanted to see complete was the *Romanian Linguistic Atlas*. Since immediately after the establishment of the new research institute no funds were available to conduct direct linguistic surveys for the linguistic atlas, nor were there any trained researchers who could effectively implement the research tool or who could perform field surveys, a correspondence survey was launched, the first questionnaire being sent in 1922. This was a method that had been previously used both in Romania and in the West in order to collect linguistic data. Such investigations were initiated by G. Wenker in 1876 [Rabanus, Lameli *et al.*, 2003: 160]. Other indirect linguistic surveys were conducted by Hasdeu, in 1884 concerning spoken language for his dictionary, *Etymologicum Magnum Romaniae*, L. Gauchat, started in 1900 and lasted for 12 years, and F. Melcher, started in 1904 [Pop, 1926: 44–45, 51, 68–69]. Furthermore, the political context after the First World War made possible and without difficulty the collection of linguistic material from all over Romania. Moreover, the argument that the social and the economic changes would cause strong linguistic changes, especially in terms of language uniformity, determined Pușcariu to immediately start language surveys through correspondence.

The material that was entered in the file of the *Romanian Language Dictionary* was extracted from the answers to the questionnaires, but at the time of writing it often happens that the lexicographer has to resort to the main source for checks. Given that the collection of answers to the *Romanian Language Museum's Questionnaires* was accessible only in physical format there were several problems inherent in such an action: on the one hand, the operation was energy and time-consuming, and on the other hand these documents are already old, some of them approaching 100 years of age and were subject to continuous physical wear. In order to address these access needs and in order to make browsing these resources an easier process, but also as a response to the urgency of digitization which could in the future involve the intention to capitalize the collection content, we set out to scan the entire archive collection and to record the documents in a digital library that would facilitate access to documents. This paper aims to show how the Calibre digital library has been customized so that it allows the efficient management of this collection but also how it can be used as a quantitative research tool on the collection of documents.

2 *The Collection of Answers to Romanian Language Museum's Questionnaires*

The documents we refer to were obtained in response to the eight theme questionnaires initially distributed in locations of Greater Romania during the above-mentioned period through school inspectorates and archdioceses and subsequently sent to those who had shown interest in becoming corresponding members of the Romanian Language Museum. The questionnaires are those listed below along with data on their size and number of responses received, according to the *Register of Correspondents for Atlas*, a document kept at the National Archives of Romania – Cluj County Service (Collection 225, The Romanian Language Museum, inventory no. 218, Register 19. *The register of entered documents of the Romanian Linguistic Atlas 1927–1938*):

1. *Questionnaire I. The Horse* [1922] was made up of 142 questions and received 670 answers.

2. *Questionnaire II. The House* [1926] was made up of 489 questions and received 439 answers.

3. *Questionnaire III. The thread* [1929] was made up of 150 questions and an Appendix to the questionnaire and 135 answers were recorded for it.

4. *Questionnaire IV. Place name and person name* [1930] was made up of 164 questions and an Appendix to the questionnaire. 182 responses were received.

5. *Questionnaire V. Sheep, Shepherding and Milk Preparation* [1931] was made up of 182 questions and an Appendix to the questionnaire and were recorded 130 answers from corresponding members.

6. *Questionnaire VI. Beekeeping* [1933] was made up of 190 questions and an Appendix to the questionnaire and recorded 45 answers.

7. *Questionnaire VII. Musical instruments* [1935] was made up of 168 questions and 49 answers were received out of which, unfortunately, none is preserved today [Comșulea, 2015: 52].

8. *Questionnaire VIII. Food and Drink* [1937] was made up of 113 questions and recorded 68 answers.

We would like to point out that in our approach to digital archiving we did not take into account the few files from the Appendices to the questionnaires that are still preserved today along with some answers. In most cases these documents consist of sheets of 8 cm / 5 cm in size on which were written words covering the topic of the questionnaire and about which correspondents were asked to specify whether they were used in their villages and what their meaning there was. Most of the useful material in these documents was extracted and was entered in the Dictionary file, which is, in fact, the purpose for which the questionnaire Appendix was used. This way of working with removable sheets had been inspired by Louis Gauchat's method. He would send forms to the correspondents so that they would complete their answers. That way the material was easy to use for the *Glossaire des patois de la Suisse romande* that he was preparing [Berindei, 2020: 175].

Registered in the order of their entry into the Romanian Language Museum, the documents were subsequently ordered according to other criteria. First of all, it should be noted that the answers that provided incomplete or no information about the informant or about the place of origin of the answer were kept separately, without being entered in the archive. This collection was set up, for the most part, according to the principles of linguistic geography in a network of points the numbering of which started from the same area as the survey points for the project on the Romanian Linguistic Atlas. The answers to the first two questionnaires are numbered consecutively, from the first to the last. It should be noted that in case of multiple answers from a village, the former would receive an additional identification symbol, either by numbers or by letters. The answers to the following two questionnaires, Questionnaire III. Thread and Questionnaire IV. Place name and person name follow the same principle of the network of cartographic points, but the numbering is no longer consecutive. It follows exactly the numbering of the survey points from the Romanian Linguistic Atlas network. The answers to the other three questionnaires within the collection are numbered in alphabetical order of the name of the place of origin. The small number of answers received and the idea of processing the material in linguistic maps was abandoned, which most likely determined the change of the principle according to which the material was ordered in the collection.

If in the bibliography of the Dictionary the identification of the manuscript was made based on the document number within the cartographic network or the serial

number that it had received according to another criterion within the corpus of answers to a questionnaire, in the digital library labelling) was made taking into account some defining elements: the name of the collection of documents, the number of the questionnaire, the title of the questionnaire in an abbreviated but clear form and the identification number of the answer to the questionnaire, as follows: MLR_CHEST_II_CASA_07, MLR_CHEST_III_FIRUL_962, MLR_VIII_MANCARI_017.

3 *Digitization and Management of the Collection Material Using the Calibre E-book Library Tools*

The collection is made up of documents that have the potential for frequent consultation. Yet, on the one hand, the physical consultation of manuscripts was time consuming and on the other hand made the materials were vulnerable to physical wear. For this reason, there were a series of measures imposed which were meant to limit the physical wear of documents. Thus, digital copies of manuscripts by means of scanning them were made and copies of the answers in .pdf and .tiff format were obtained, at the maximum quality allowed by the CZUR ET16 Plus scanning station.

In their documentation efforts lexicographers often need to identify the origin of all the manuscripts by the same correspondent or the identifications of all answers to questionnaires from a particular location. Thus, we have opted for managing the collection of documents in digital version via Calibre Library in order to make access to these documents easier and more efficient for the researchers. This is „an open-source e-book management tool. Simply put, calibre allows you to organize your e-book library, convert e-books to various formats, and interact with your e-book reader, all in an intuitive and friendly manner” [Schember, 2010–2014].

Calibre Library is customized with several columns which are specific to book management, in which you can enter the metadata corresponding to each title added to the library. The columns are Title, Languages, Tags, Size, Rating, Date, Series, Publisher, Published, Authors, Modified. These are not user-defined columns and cannot be modified or deleted by the person configuring the database, but one can choose not to make them visible on the layout of the columns in the Calibre library book list, as it can be seen in the dialog box below.

However, Calibre Library allows the addition of new columns for entering other types of metadata, which made us choose this tool for managing the manuscripts we are dealing with. As it can be seen in the image below, new categories of metadata can be added in the columns, which can be visible or hidden in the user interface, in unlimited numbers, thus allowing the efficient use of the platform for handwritten documents, such as those in our collection.

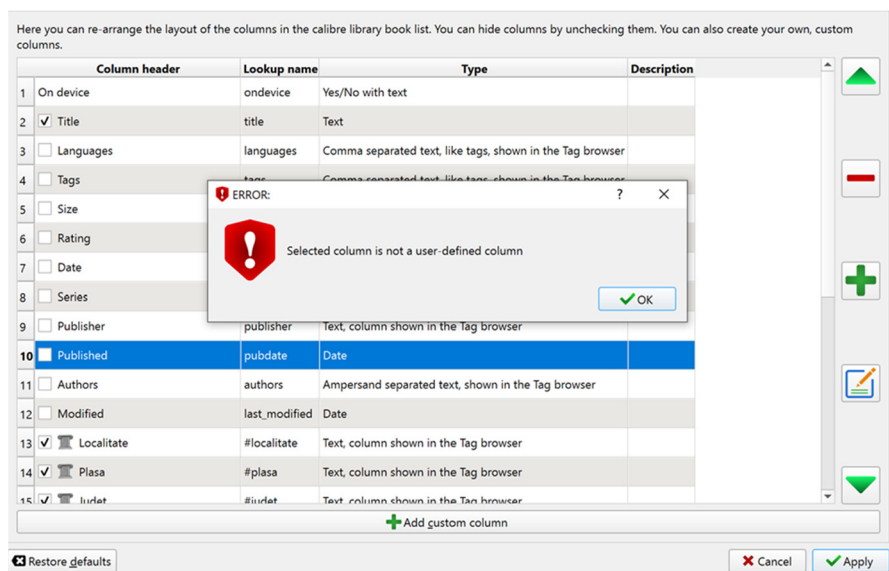


Figure 1 Box dialog: re-arranging the layout of the columns in the Calibre library

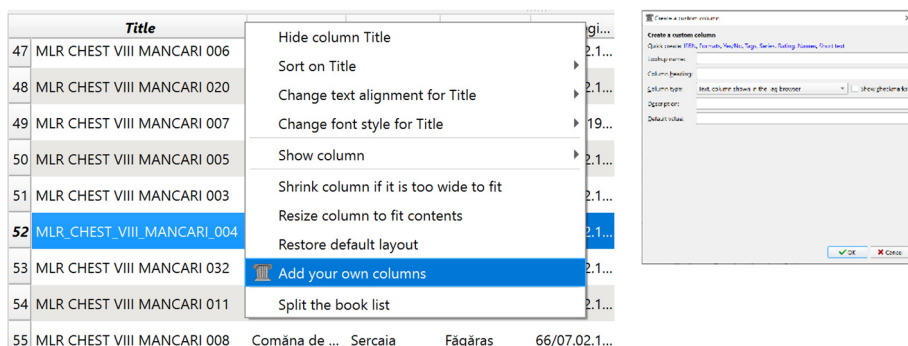


Figure 2 Adding new categories of metadata

Searching in Calibre Library is also very simple using the search box dialog: „Just type a few key words into the Search field. Try the author, title, series, or anything else from the e-book’s metadata. E-books matching your search terms are shown as filtered results. The other e-books are still in your library, but they won’t be shown if the search expression doesn’t find them” [Schember, 2010–2014].

4 *Adapting the Calibre Library for the Management of the Manuscripts in the Collection of Answers to the Romanian Language Museum's Questionnaires*

As we have already pointed out, Calibre offers the user the possibility to configure the database from the first use, depending on the user's preferences. In addition to user interface configuration options, Calibre also supports multiple database customization options by adding or by deleting metadata categories.

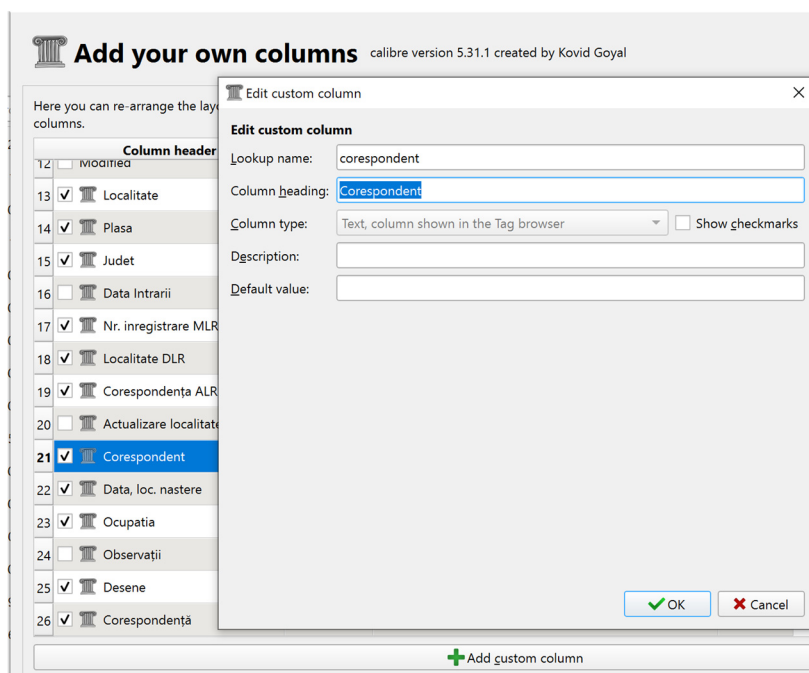


Figure 3 Adding the column Correspondent in the Calibre Library

For the use of the Calibre digital library in managing the documents of the Romanian Language Museum's Questionnaire collection, the database was customized by entering the following metadata categories: geographical location of the answer, correspondent's name, date and place of birth, correspondent's occupation, date of entry of the manuscript in the Romanian Language Museum and the registration number, information about the drawings and about the correspondence contained. Another category of metadata is the one regarding the correspondence, if any, between the location from which the answer was sent and an investigation point from the Romanian Linguistic Atlas network. The last three categories of metadata were not initially included in the project; they were

additional and provided new research tracks on the material. Therefore, they are not visible in the user interface and are not part of the metadata that facilitates lexicographers' access to the archive collection.

The metadata categories related to the geographical location of the answer has six columns named as follows: Location, District, County, Location (according to) DLR, Updated location, Cartographic point ALR. The first three of them represent the information contained on the Individual sheet accompanying the answers or on other documents that helped with their identification (document cover or last page, Correspondence Register for the Atlas). The fourth category of metadata refers to the location of the answer according to the administrative division at the time of the publication of the information in the bibliography of the Romanian Language Dictionary, in 1965. The following category of metadata refers to the location update according to the current administrative division of the country and is connected to the map network of the Romanian Linguistic Atlas.

5 Conclusions

As shown above, the use of this digital tool is a way of making accessible the material resulting from the indirect linguistic survey carried out at the *Romanian Language Museum* in Cluj-Napoca during the period between the two World Wars. It is a very useful internal work tool which is accessible within the Romanian Academy and is conceived to offer researchers the opportunity to consult the material avoiding time and energy consuming actions that browsing the material in physical format would normally involve. Additionally, it is also a means to protect manuscripts which are nearly a century old. The design of this tool was all the more necessary as the collection of answers to the *Romanian Language Museum's Questionnaires* is a lexicographical source used in the project *Romanian Language Dictionary*, as it was introduced in its bibliography. Furthermore, the need to check information is frequent, which makes the creation of this digital tool even more useful.

However, this digital tool aims at each document as a whole, it is accessible in .pdf format and the metadata naturally refers to the document as a whole. Although it is a very useful tool and its implementation was a real necessity, it does not manage the true richness of the collection of documents that is revealed by the content of each document, i.e., the linguistic data that can be retrieved from the answers to the questions asked.

This small step towards digitization and digital archiving opens the possibility to future approaches to the archive collection, which will capitalize on the content of each manuscript in all its details. Excerpts of the answers to each question, correlated with the question and benefiting from annotations, as well as the establishment of a digital glossary of terms covering the entire collection of documents may represent future research tracks.

Moreover, the identification of the correlations between the survey points of the *Romanian Linguistic Atlas* and the points from which the answers to the questionnaires were received involves a first step towards the interconnection of the two collections and the entry of this collection in the scientific circuit of linguistic geography as an auxiliary element. Future processing of data, similar to the one in which the collection of answers to Wenker's indirect survey was capitalized, which helped set up *The Digital Wenker Atlas* (www.diwa.info) [Herrgen, 2010] may open new perspectives on this collection of documents.

Acknowledgements

This work was supported by Romanian Academy Grants within the project Digital Library of the Romanian Language Museum's Questionnaires' Corpus. Integrated System of Metadata Management (BIBLIO – MLRom), a research grant supported by Recurring Donor Fond, available to the Romanian Academy and managed by „PATRIMONIUM” Foundation, GAR-UM-2019-I-1.5-5, contract number 359c/15.10.2019.

References

- Berindei, C.–M. (2019). Construcția instituțională a Muzeului Limbii Române ca loc al memoriei. In *Caietele Sextil Pușcariu*, IV: 71 – 81.
- Berindei, C.–M. (2020). Ancheta lingvistică indirectă a Muzeului Limbii Române – câteva considerații asupra metodei de cercetare. *Dacoromania*, s. n., 25(2). 173–188.
- Chestionarul I. Calul* (1922). Cluj, Muzeul Limbii Române, Tipografia Ardealul.
- Chestionarul II. Casa* (1926). Sibiu, Muzeul Limbii Române, Dacia Traiană.
- Chestionarul III. Firul* (1929). Cluj, Muzeul Limbii Române, Tipografia Ardealul.
- Chestionarul IV. Nume de loc și nume de persoană* (1930). Cluj, Muzeul Limbii Române, Tipografia Ardealul.
- Chestionarul V. Stâna, păstoritul și prepararea laptelui* (1931). Cluj, Muzeul Limbii Române, Tipografia Ardealul.
- Chestionarul VI. Stupăritul* (1933). Cluj, Muzeul Limbii Române, Tipografia Ardealul.
- Chestionarul VII. Instrumente muzicale* (1935). Cluj, Muzeul Limbii Române, Cartea Românească.
- Chestionarul VIII. Mâncări și băuturi* (1937). Cluj, Muzeul Limbii Române, Tipografia Ardealul.
- Comșulea, E. (2015). Chestionarele Muzeului Limbii Române – premise de cercetare. *Philologica Jassyensia*, IX, 1 (21): 47–56.
- Haja, G., Dănilă, E., Forăscu, C., Aldea, B.–M. (2005). *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea*, Iași, Editura Alfa.
- Herrgen, J. (2010). The digital Wenker Atlas (www.diwa.info): an online research tool for modern dialectology, *Dialectologia: revista electronică*, Special issue, I (2010): 89–95.
- Schember, J. (2010–2014). *Calibre Quick Start Guide*. Fourth Edition. Licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. © 2010-2014 John Schember.

- Pop, S. (1926). *Buts et méthodes des enquêtes dialectales*. Extrait des „Mélanges de l'École Roumaine en France”, II-e partie, Paris, J. Gamber.
- Pușcariu, S. (1921). Muzeul limbei române. *Dacoromania*, 1: 1–8.
- Rabanus, S., Lameli, A., Schmidt, J. E. (2002). La geografia linguistica tedesca e la Scuola di Marburg, *Bollettino dell'Atlante Linguistico Italiano*, III Serie, 26: 159–184.

CHAPTER 2

CURATION OF DIGITAL HERITAGE

The Role of the Philologist in the Project Artificial Intelligence Models (Deep Learning) Applied in the Analysis of Old Romanian Language (DeLORo)

GABRIELA HAJA and ELENA ISABELLE TAMBA

A. Philippide Institute of Romanian Philology. Romanian Academy, Iasi Branch
{gabi.haja, isabelle.tamba}@acadiasi.ro

Abstract

This paper presents the project Artificial Intelligence Models (Deep Learning) Applied in the Analysis of Old Romanian Language (DeLORo) from the perspective of the philologists involved. The main goal of the project is to build a technology that would allow the automatic transposition of Old Romanian semi-uncial manuscripts and prints from the Cyrillic alphabet or the mixed alphabet – Cyrillic and Latin. In addition to the main objective regarding the technology of transliteration of Cyrillic texts, specific objectives were outlined such as the automatic identification of Romanian language models (lexicology, morphology / lexical, morphological) for each half century of the considered period (the 16th century until 1862). The philologists' activity in the project is determined by the achievement of these objectives. Therefore, the composition of the corpus of scanned texts (with or without published critical editions), the manner of manual annotation of text samples (transliteration vs. transcription) are intended to facilitate the training of the word recognition technology, to transpose them in Latin alphabet and to create the premises for the material provided by philologists (by annotation and transliteration / transcription) to ensure valid results for the automatic generation of language models.

1 Introduction

DeLORo is an experimental demonstrative project, which aims to create technology for recognizing and transliterating¹ Romanian texts, both printed and manuscripts with Cyrillic or mixed alphabet – Cyrillic and Latin.

Unlike similar projects – like Abbyy FineReader or Transkribus – which only consider the optical recognition and reproduction of characters in an image, the DeLORo project aims to transpose the scanned Cyrillic text into the Latin alphabet. Another difference compared to projects such as those previously mentioned is that the results obtained through DeLORo will be available, open

¹ The transliteration faithfully reproduces in the Latin alphabet the original written in Cyrillic.

access, to specialists (philologists, historians, etc.), and to the general public, interested in old Romanian documents alike. In the project's team 4 experienced philologists are included.

2 *DeLORo Corpus*

The DeLORo corpus includes 60 scanned documents, selected in accordance with several criteria: temporal criterion, spatial criterion, text type (print / manuscript), functional style, language used, difficulty of the text and, implicitly, of the annotation. For creating this corpus, philologists took into account different types of texts, grouped according to these criteria as follows:

- by content: religious and secular (scientific text, literary text, journalistic text)
- by form: semi-uncial manuscripts and prints (with Cyrillic and mixed alphabets – Cyrillic and Latin)
- by geographical distribution: as far as possible, from all historical regions
- by the difficulty of annotation (quality of the original; type of writing; type of alphabet, etc.)

The main source of the text collection is the Library of the Academy of Bucharest, where they were digitized. If some titles had not been included on the list of scanned books, they were recovered by the efforts of Mrs. Gabriela Dumitrescu, head of the Manuscript Service, Rare Books. We present in Figure 1 some fragments from the table made up by the philologists, at the beginning of the project.

In this table we can find information about the chronology, the author(s) / the editor(s), the title of the work, the place where each text was published or written, if it is a print or a manuscript, the content – religious or secular (scientific text, literary text, journalistic text, etc.), the language and the alphabet (Cyrillic and mixed alphabets – Cyrillic and Latin), the difficulty level, number of pages, etc.

An	Autor	Titlu	Loc	Tipar/ms.	Stil	Limba	Dificultate relativă	Cota BAR	Transcris (editat critic / transcris)	Nr pagini	Metadate XML BAR
1500/1520	(?)	<i>Psaltirea Hurmuzaki</i>	Maramureș (?)	uncial	beletristic (text biblic)	română	3	MS. rom. 3077	DA	249 p.	DA
1559/1564	Coresi	<i>Catehism</i>	Brașov (Trans S = Munt)	tipăritură	bisericesc	română	1	MS. ROM. 5032	DA	21 p.	NU
1563	Coresi	<i>Apostol</i>	Brașov (Trans S = Munt)	tipăritură	bisericesc (text biblic)	română	1	CRV 12	DA	618 p.	DA

cca 1658-1661	Kigalas (autorul originalul ngr.)	<i>Hronograf den începutul lumii Ms. 3517</i>	Mold.	ms. uncial	științific, beletristic traducere	română	2	MS. ROM. 3517	DA	608 p.	NU
1700	Radovici, Gheorghe (tipograf, pref.) Filotei Svetagorețul (trad.)	<i>Învățăture creștinești foarte de folos, acuma întâi scoase de pre limba grecească, pre limbă rumânească...</i>	Snagov Munt.	Tipar negru; 21 de rânduri pe pagină	bisericesc	română	1	CRV 118 De făcut		225 imagini	NU

1799	Iorgovici, Paul	<i>Observații de limba rumânească. Gramatica rumânească</i>	Buda	Tipar negru; 20-23 rânduri pe pagină	științific	română	1	CRV 621	DA	116 imagini	DA
1806	Micu-Klein, Samuil	<i>Istoria, lucrurile și întâmplările românilor pre scurt așezată și din mulți vechi și noi scriitori culeasă și scrisă</i>	Buda	Tipar negru în pagină cu chenar, 28 de rânduri pe pagină	științific	română	1	CRV 695		92 imagini	NU

Figure 1 Excerpt from the table containing the corpus entries

In the following, we will present the different types of texts that we consider in the project, depending on the degree of difficulty of the annotation (Figures 2-8).

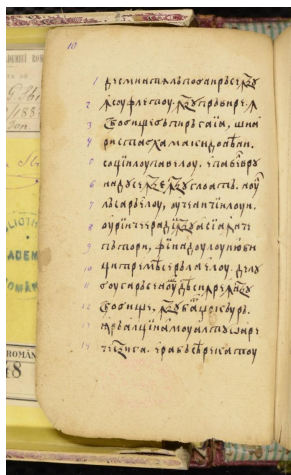


Figure 2 Semi-uncial manuscript
(extract from *Codicele Voroneșean*,
1563–1583)

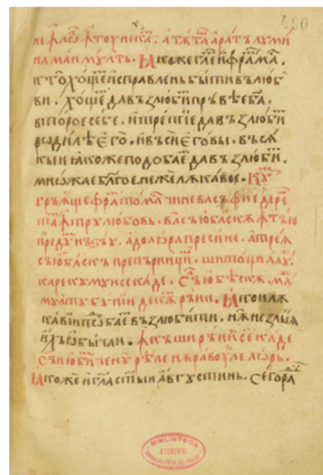


Figure 3 Semi-uncial manuscript
Slavo-Romanian, with a medium level
of difficulty of annotation (extract
from *Floarea darurilor*, Putna,
1592/1604)

For the texts from the 16th – 17th centuries, we benefit from the results obtained from the Old Text project from Iași (<https://textvechi.ro>). The images from this period are annotated (at line and character level), using the interpretative transcripts available.

For the texts from the following centuries the situation is different: most of the selected documents were either not (re)edited or were published before 1990 and we do not have the transcripts in an editable electronic format, without errors. For the last ones, we opted for transliteration instead of interpretative transcription, so as to provide the DeLORo technology with a model that involves as few situations of ambiguity as possible. We tried, therefore, to render the original text as accurately as possible, without lowering the overwritten characters or filling in the abbreviated words. Also, to eliminate ambiguities and facilitate the recognition and equivalence of words, we opted for spelling with *î*. We needed to make decisions on issues that traditional philology tacitly solves: marking the pause between words, respecting the current spelling in the case of spellings, marking at the end of the line the division into syllables, etc. We opted for the transcription with pauses between words, rendering the spellings according to the current norms, with the marking at the end of the line of the division into syllables, etc.

The results obtained in two other research projects are also used to create textual corpora and language models in our project: *Monumenta linguae Dacoromanorum*. Bible 1688 (MLD) and SOLIROM¹. The MLD database, made by researchers from Iasi in the project of critical re-editing of the Bucharest Bible in parallel with two other translations from the same century, ms. 45 and ms. 4389, includes the complete morphological annotation of the 1688 Bible text, with the establishment of the lemmas by the publishers (starting with the volume printed in 2008² until the end of the edition in 2015³). The collection of inflected forms (ca 400,000) correlated with the corresponding lemmas from the MLD base is a very useful resource for project's activities.

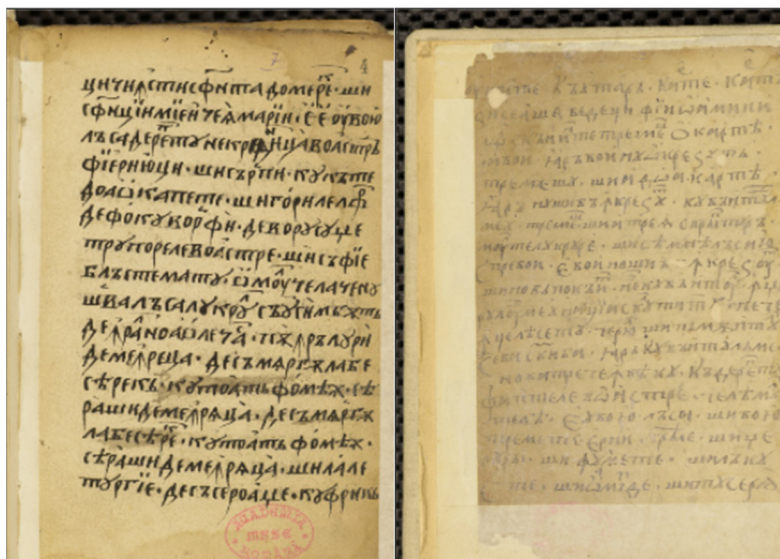


Figure 4 Semi-uncial manuscript, with a high level of difficulty of annotation (extracts from *Codicele Bratului*, 1559–1560) – (opaque) spots, oblique or curved lines, indecipherable words, etc.

Therefore, the final database will contain a collection of images annotated with linguistic and graphical data and metadata, which will be a resource for future research, which may lead to the optimization of the results expected in the DeLORo project.

¹ <https://solirom.ro/>

² <https://biblia1688.solirom.ro/7/>

³ http://consilr.info.uaic.ro/~mld/monumenta/MLD_VIII/start.html

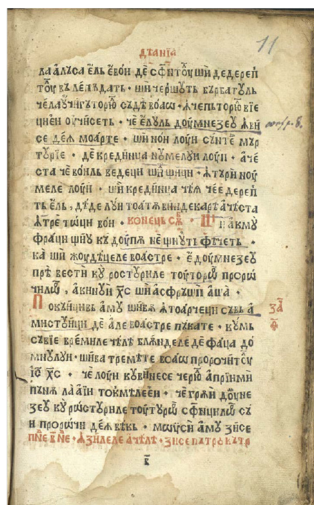


Figure 5 Print with a medium level of difficulty of annotation due to spots, marginal handwritten notes, handmade underlining, etc. (extract from Coresi, *Apostol românesc*, Braşov, 1563).



Figure 6 Print with a high level of difficulty of annotation due to overlapping print (extract from *Catavasiariu tîlmăcit pre limba rumânească cu multe trebuincioase cântări ce să cântă preste tot anul, care s-au tipărit acum...*, Bucureşti, 1742)

3 OOCIAT Interface for Annotation (Online Old Cyrillic Image Annotation Tool)

The OOCIAT⁴ interface allows philologists to annotate the image and transcribe the related text. It is a working tool that facilitates remote access of users (annotators) to the DeLORo database, which includes the collection of documents in image format accompanied by metadata and data resulting from their manual annotation. Annotated graphic objects are presented in Table 1.

Basically, the annotation operation by philologists consists in drawing a border around the annotated object, establishing the type of object, transliterating / transcribing the content in the case of primary and secondary objects. This way, for each annotated object the spatial coordinates within the image / page are assigned, some of them being decoded in writing. For graphic objects, their meaning is given by the selected type.

⁴ OOCIAT interface was designed and made by PhD student Cristian Pădurariu.

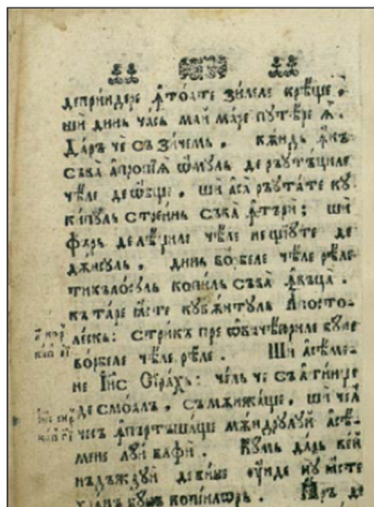


Figure 7 Print with a medium level of difficulty of annotation, due to obliquely printed lines (extract from *Bucvar, Iași, 1755*)

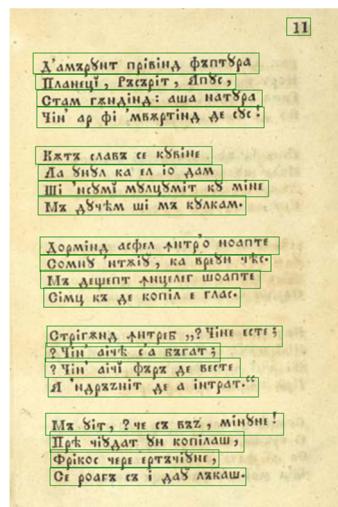


Figure 8 Print in mixed alphabet (extract from *Ioan Văcărescu, Poezii alese [București], 1830*)

4 DeLORo Golden Corpus

Table 1 Types of objects annotated in DeLORo project

Primary objects	Secondary objects	Graphic objects
<title>	<modifier ⁵ >	<frontispiece>
<column> => L,R	< marginal text>	<ornament>
<line>	<interlinear text ⁶ >	
<character>	<capital letter>	
	<marginal reference ⁷ >	
	<interlinear reference ⁸ >	

Many of the automatic natural language processing programs are trained to perform certain functions based on "golden corpora", made manually by linguists,

⁵ Modifiers signal additional information about the primary objects – accent and title.

⁶ The interlinear text is represented by any character/ group of characters that is overwritten.

⁷ The marginal reference represents the repeated sign(s) that announces a marginal textual notation.

⁸ The interlinear reference represents the sign(s) that announce in the text, in an overwritten position, a marginal textual notation.

in a machine-accessible language (most often, XML). In the annotations that are made in this language, the linguistic information is labelled with additional information, of various categories and levels. In the training stage, it is important that the language specialists correct the errors generated by the recognition programs thus trained, the corrections sometimes helping improve the programs' behaviour, if new input data is added.

For DeLORo technology, it was decided to compile a gold corpus containing a minimum of 120 annotated pages for each half century. It was also decided that of these, 60 would be pages with minimal difficulty, 40 – pages with medium difficulty, 20 – pages with high difficulty.

Print in Cyrillic alphabet

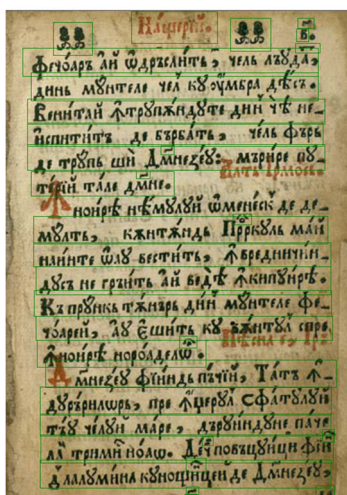


Figure 9 Page with medium difficulty. Catavasiariu tălmăcit pre limba rumânească cu multe trebuincioase cântări ce să cântă preste tot anul, care s-au tipărit acum..., București, 1742

Ornament:null Ornament:null Titlu:Nașterii.
 Rând:Venit-ai intrupindu-te din cea ne-
 Rând:dinș muntele cel cu umbra deasă.
 Rând:fecioară ai odrăslitș, celș lăuda.
 Rând:de trupș și Dmnezu: mărire pu-
 Rând:nainte o au vestitș, invrednicin- Rând:terii tale, Dmne.
 Rând:ispitită de bărbatș, celș fără
 Rând:cioarei, au eșitș cuvintul spre
 Rând:durărilorș, pre ingerul sfatului
 Rând:tău celui mare, dăruindu-ne pace
 Rând:Dmnezeu fiindș păcii, Tată in-
 Rând:du-să negrăitș ai vedea închipuirea.
 Rând:î-a trimis noao. Dec povățuiți fii
 Rând:innoirea noroadelo.
 Rând:Că pruncș tinărș din muntele fe- Rând:de
 Rând:d la lumina cunoștiei de Dmnezeu, Literă:2, Literă:t
 Literă:o Literă:r Literă:n Literă:2 Literă:i Literă:n
 Modifier: Modifier: Modifier: Modifier:

In creating this corpus we encountered the following problematic situations:

- for the first half of the 16th century very few documents have been preserved;
- the presence of bilingual (Slavo-Romanian, Romanian-Slavic) or multilingual (see *Codex Sturdzanus*) texts; in these documents, only the Romanian text is annotated;

- semi-uncial manuscripts are specific to the 16th and 17th centuries; starting with the 18th century, cursive writing is used, which computer specialists do not consider at this stage; for this period, we will annotate only prints;
- 19th century documents present difficulties arising from the use of mixed alphabets (Cyrillic and Latin).

Print in mixed alphabet (Cyrillic and Latin)

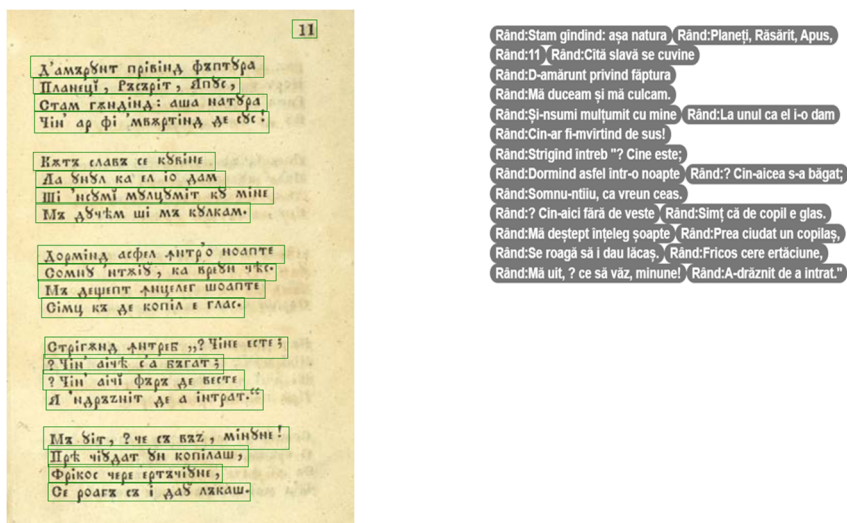


Figure 10 Page with minimal difficulty. Ioan Văcărescu, *Poezii alese* [București], 1830

5 Conclusions

The contribution of philologists in the project is important, because the accuracy of the results obtained by DeLORo technology depends to a large extent on the volume and quality of the activity of manual annotation of the corpus for training. However, the annotation process is a chronophagous one.

From our perspective, the fact that the technology created within the project will be made available to researchers can be useful, given that similar services, already developed, are expensive. In the current phase, we expect that the results of the automatic transcription will not be impeccable, but will facilitate future editions of Old Romanian texts, and at a later stage, the DeLORo technology can be improved as new corrected data will accumulate. The process of accumulating training data is similar to rolling a snowball, because the documents decoded by

this technology will have to be corrected by human experts, in a shorter time, of course, than would have required their full annotation, then adding this to the training corpus.

Even in these conditions, by realizing this technology, the premises for a higher level digitalization of the collections of Old Romanian documents are created. It's also important to offer this technology to the public.

Acknowledgements

This paper presents some aspects of the project *Artificial Intelligence Models (Deep Learning) Applied in the Analysis of Old Romanian Language (DeLORo – Deep Learning for Old Romanian)*, UEFISCDI – PN-III-P2-2.1-PED-2019-3952, no. 400PED; coordinator: dr. Dan Cristea m. c. A.R.) (2020–2022).

Data Structure and Acquisition in DeLORo – a Technology for Deciphering Old Cyrillic-Romanian Documents

DAN CRISTEA^{1,2}, CRISTIAN PĂDURARIU^{1,2}, PETRU REBEJA²,
ANDREI SCUTELNICU^{1,2}, and MIHAELA ONOFREI^{1,2}

*¹Institute for Computer Science, Iași branch of the Romanian Academy,
2 Codrescu St.*

*²Faculty of Computer Science, Alexandru Ioan Cuza University of Iași,
16 Berthelot St.*

dan.cristea@acadiasi.ro, {cristian.padurariu94,
petru.rebeja, andreiscutelnicu,
mihaela.plamada.onofrei}@gmail.com

Abstract

DeLORo (Deep Learning for Old Romanian) is a project intended to build a technology capable of deciphering old printed and uncial Cyrillic Romanian documents and transliterate them into the Latin script. In this paper we concentrate on the processes of organisation and of acquisition of the primary data necessary in the process of training the deep learning recognition technology. After a brief overview of similar enterprises, we compare our approach with others. Then, we present in some details the structure of DeLORo's data repository, which includes: images of scanned pages, annotations operated over them, and alignments between annotated objects in the images and sequences of decoded Latin characters. The types of annotated objects, as facilitated by an interface, and the accessing details in the resulting database are then presented. We overview the data collected so far and argue for a strategy that must accelerate the process of data acquisition and the exploitation of documents entirely transcribed before the beginning of the project.

1 Introduction

DeLORo (Deep Learning for Old Romanian) is a project¹ that has the goal of building a technology capable of deciphering old printed and uncial Cyrillic-Romanian documents and transliterate them into the Latin script. The project is motivated by the great interest in preservation of the Romanian written cultural heritage and it is meant to enhance and speed up the browsing through and the reading of old Romanian documents by specialists, while also opening them to the attention of today's non-expert readers.

¹ Financed by the Romanian Government and running between October 2020 and October 2022.

The resulting technology should be capable to classify different types of visual objects in the scanned pages of the original documents used for training and evaluation and to interpret their non-graphical textual content, by converting the original Cyrillic alphabet into its Latin equivalent, respecting the contextual interpretation for those characters that may have multiple readings. The classification and interpretation operations are based on deep learning (DL) models. In this paper we describe the organisation of the database and the process of acquisition of primary data necessary in the process of training the DL-based object identification (OI) and character recognition (CR) technologies.

The Cyrillic alphabet was used to write texts in the Slavonic language (Palaeo-Slavonic) since the 9th century, then followed by the Medio-Slavonic. For the Romanian language it was used between the 16th and the 19th centuries. A unique Cyrillic alphabet has circulated on the territories of historical Romania, with slight variations individualized by differences in graphemes or their phonetic values. This situation changed in 1828, when the philosopher and linguist Ion Heliade Rădulescu turned the Romanian Cyrillic alphabet entirely into a phonological system [Gheție and Mareș, 1985]. The formal adoption of the Latin alphabet, replacing the Cyrillic, was imposed by Alexandru Ioan Cuza in 1862, after a transition period that lasted several decades. An inventory of 1968 works (Slavonic, mixed, and Romanian – the most numerous) written in Cyrillic in the interval 1508-1830 is reported by Bianu *et al.* [1903-1944]², while Căndea [2011; 2012; 2014] makes an inventory of documents held in foreign libraries. Most of these books receive an extremely limited exposure, because, apart from a small number of specialists working on paleo-linguistics, which can access rare texts in libraries, they are not accessible to the large public. Moreover, the modern editing houses are interested to print and offer to the book market only books authored in modern times, extremely few older than the middle of the XIXth century. Rare texts, if ever, are printed only as philological editions, in exquisite printing conditions, therefore usually at high prices, in which pictures reproduce fragments of the originals – see, for instance [Dumitrescu, 2017; 2018] –, mainly because of the lack of a technology able to interpret them in Latin Romanian and to acquire typographical edited versions.

The need to open access to this rich but less studied literature made out of old and rare Romanian texts, not only in their image format, but also in editable forms, motivates our research interest. As such, the DeLORo project addresses more categories of users:

- the computer scientists developing artificial intelligence and image processing methods with applications in computational linguistics, computational lexicography, digital humanities, interested to find ready-made processing tools and annotated training data;

² See also: <https://biblacad.ro/rarom.html>

- the linguist interested to decipher and study old Romanian documents, interested to operate indexing, searching, browsing and transcription mechanism on scanned textual sources;
- the historian, geographer, philosopher, legist, astronomer, theologian, reviewer, therefore to all categories of specialists in domains the old texts themselves belong to, which are interested to read those texts, but are discouraged by the impossibility to correctly decipher them, because they often lack linguistic skills specifically addressing old Romanian writings;
- the librarian, interested to open to the large public online indexed access into the content of old Romanian texts, thus also preserving the precious original paper editions out of a direct exposure in the hands of readers;
- finally, the editor, interested in quickly editing old texts in modern editions, wishing to print and sell rare books in page layouts which reflect their own creative views.

The type of processing proposed over original scanned versions of old books will convert the images of pages into an editable electronic format, out of which the users will be able to easily select and edit, as they prefer, the original content. DeLORo will thus become an environment that, put in the hands of all these categories of users, will bring them a completely new and exciting opportunity of browsing the written cultural heritage, opening their universe towards a rather hidden part of Romanian culture. Moreover, the set of algorithms developed in the project and the collection of training data will be deposited on a platform that can be freely accessed by expert users for research and education purposes³.

The paper is organised as follows. Section 2 briefly describes similar studies and technologies. Section 3 presents the way data is structured in DeLORo's repository, called ROCC (Romanian Old Cyrillic Corpus). Section 4 shows more ways for data acquisition and some statistics in the middle of the project's development, including a strategy thought to accelerate the process of data acquisition. Section 5 describes the used techniques for searching information in historical documents, as well as ROCC statistics. The paper concludes with plans about the future work and some discussions.

2 *Related Work*

From the linguistic point of view, in [Gafton, 2016] there are formulated a number of observations concerning the transcription of old Romanian texts, written in Cyrillic, which are considered for being applied in transliterating our collection of old texts. Gîfu [2015; 2016] describes a chain of linguistic processing applied to a collection of Romanian texts collected over the territories of Romania and

³ <http://lsplr.iit.academiaromana-is.ro/>

Bessarabia to measure lexical, syntactic and semantic similarities between the variants of Romanian language used in the two provinces. Plămadă-Onofrei [2018] edited the first Geography textbook written in Romanian, transcribing it from Cyrillic Romanian into Latin Romanian alphabet. As a philological operation, the editing of this textbook, *De obște gheografia*, meant to reveal the detailed linguistic study, highlighting the phonetic, lexical, morphological and syntactic peculiarities of this text.

At the level of processing, DeLORo is rooted in an enterprise aiming to accurately estimate the contribution of historical events on the evolution of the Romanian language [Cristea *et al.*, 2018]. Later, Pădurariu [2019] makes a thorough investigation of the state of the art in image segmentation and object focusing methods, his findings being applied in our project. The approaches at character level proved to influence the development of lexical clustering models. For instance, Ionescu *et al.* [2016] show that the methods that work at the character level are very effective in tasks like native language identification, while Năstase and Popescu [2009] show their effectiveness in lexical tasks like grammatical gender prediction. Spectral clustering, successfully applied to word sense disambiguation [Popescu and Hristea, 2011], will also be applied here. Many other theoretical and practical achievements of members of DeLORo contribute to the scientific and implementational knowledge which make the basis of our project, among which: the national projects eDTLR [Haja and Cristea, 2010], MappingBooks [Cristea *et al.*, 2018a] and CoRoLa [Cristea *et al.*, 2017].

A similar project which focused on digitisation and preservation of old texts written in Dutch, German, English, French, Spanish, Polish, Czech, Slovene and Bulgarian is the project IMPACT4, funded by the European Commission. This project aimed to improve the access to historical texts published between the XVth and XIXth century. For converting digitised images into searchable and editable full-text, they used a commercial OCR engine, developed by ABBYY, not an in-house algorithm. At the end, the project reported an accuracy rate of 90%-word recognition.

3 Data Organisation

Since the data involved in DeLORo is heterogeneous in nature and needs to cater for different needs (load of images, CRUD⁵ on the annotations contributed by humans, data export for training DL models, etc.) we adopted a hybrid approach for data-storage, namely keeping images and text files on the file system of our

⁴ <http://www.impact-project.eu/>

⁵ An acronym that abstracts the four basic database operations, namely: Create, Read, Update, and Delete.

dedicated server and the rest of the data in a PostgreSQL database. Whenever the format of the data needs to be adapted for a specific use-case we pass it through a transformation pipeline⁶ which outputs the required format. Within the DeLORo project we distinguish three main types of data: (i) *imported data*, (ii) *core data*, and (iii) *exported data*. We describe them below.

3.1 Imported Data

By *imported data* we denote data that was created beyond the time-frame of the DeLORo project which, through either copying as-is to the database or additional processing is transformed into the core data. The imported data consists of three main categories: (i) scanned images of pages, (ii) metadata, and (iii) transcribed texts. Out of these three categories, only the transcribed texts are copied as-are into the core data of the project; additional processing is applied to the other two. Each imported book is either a PDF file containing scanned images of pages, or a directory containing page images in JPEG format. Furthermore, for each book, an additional XML file keeps the corresponding metadata (title, authors, creation year, publishing info, etc.). The import pipeline iterates each directory corresponding to a certain book, applying the following actions: (a) replicates the directory in the core data; (b) runs the metadata import tool⁷, which iterates over the metadata files, and updates metadata values accordingly into the core data; (c) if the imported directory contains images, the images are converted into the PNG format, then images are renamed, to ensure lexicographic ordering of pages, and copied into the core data; (d) if the imported directory contains just one PDF file, which concentrates all images of pages, then the images are extracted into the core data directory as separate files in the PNG format, with names attributed to replicate in the lexicographic order the order of pages in the originals.

3.2 Core Data

The organisation of DeLORo *core data* is described in [Cristea *et al.*, 2020] as ROCC (the Romanian Old Cyrillic Corpus). Here, a page of a document is registered by the following components: metadata attached to the whole original document and to each page (required), the page image (required), image segmentation by humans and/or machine (optional), transcribed text (optional), and image-text alignments (optional).

Furthermore, image segmentation is specific to the type of annotated objects. In the corresponding XML schema of the core data, each document is denoted by

⁶ This pipeline consists of a script that encodes the transformation and a process that, when triggered, runs the transformation script and applies additional manipulation onto the resulting data, e.g. (un)archiving, moving to a specific directory, etc. These pipelines are stored on Github at: <https://github.com/deloro-project/rocc-pipelines>.

⁷ The source code of the metadata import tool is not publicly available yet but will be published soon.

the `<pageCollection>` tag, placed at the top of the ROCC corpus. As such, a collection of pages characterizes a document, i.e., a book, a newspaper issue, generally an editorial unit that has an associated library identification code. At the level of metadata, this description combines information taken from the sites of the libraries holding old Romanian documents, with others validated and then completed in the discussions among members of the DeLORo consortium.

Where transcripts are available, the scanned images of the Cyrillic pages are presented in the core data in pairs with their Roman textual transcripts. The basic unit in image-to-text alignment is the page.

The XML schema denotes the following objects that carry a content (i.e., a sequence of Latin characters, as decoded from Cyrillic by annotators):

```
<object:Title> <!-- for catchwords, usually appearing on top of pages; -->
<object:Column> <!-- for columns (a page may contain one or two columns);-->
<object:Line> <!-- for lines of text;-->
<object:Character> <!-- for Cyrillic characters;-->
<object:Marginal> <!-- for text areas on the cuff;-->
<object:OutOfLineCharacters> <!-- for a letter or a sequence of letters placed outside of rows
    (above or below them);-->
<object:Modifier> <!-- for modifier signs placed above the words, as for instance: the sharp
    accent ('), the grave accent (`), or the tilde (~); modifiers can signal abbreviations,
    change of meaning of the character from letter to number, etc.;-->
<object:InitialLetter> <!-- for decorated letters at the beginning of the paragraph;-->
<object:ReferenceMarkOnMargin> <!-- for cross reference signs (landmark, spell, etc.)
    placed on the cuff;-->
<object:ReferenceMarkAboveLine> <!-- similar, but placed above lines.-->
```

All occurrences of these types of XML elements, with the exception of `<object:Modifier>`, include the attribute `@content`, which records their Latin transcription, done by annotators or the machine, as one or more characters in sequence⁸. Other information usually carried by these elements are as follows:

```
<graphicalObject:Frontispiece> <!-- marking frontispieces;-->
<graphicalObject:Accolade> <!-- for accolades placed on the cuff;-->
<graphicalObject:Ornament> <!-- for drawings or prints with the role of ornament or
    simulating engravings, placed either on the edge or at the top or bottom of the page;-->
<graphicalObject:Frame> <!-- for framed drawings or printed objects, usually placed
    around areas that contain text.-->
```

The graphical objects do not carry textual contents, but supplementary positioning elements could record special traits of the graphical objects (as for instance the point of the accolade, or the inner rectangle in case of a frame).

All these objects are grouped under the `<imagesOfPages>` element, which contains the sequence of page images of a document. In this area each page is recorded as `<onePageImage>` XML element having the fields:

```
@pageID<!-- automatically generated by the Platform when loading the sequence
    of page images of a document,-->
@pageImageFile <!-- the URL or the name of the file that contains
    the page in the document directory (example: tragest-lui-orest-2-pagina-39.png). -->
```

⁸ When the content is VOID, it means that the object has been identified, without any interpretation of its content.

Moreover, an element <difficultyCriteria>, manually filled in by the annotator after a visual inspection of the page, records the degree of difficulty for automatic processing of the image: if the page is damaged, if spots are obstructing the writing⁹, if there are characters appearing from the back through transparency, if there are patterns¹⁰ that overlap each other, if there is handwriting between lines, abbreviations of special signs, if the page is a palimpsest¹¹, if there are corrections on the page¹², and if there is marginal writing¹³.

3.3 *Exported Data*

Within the DeLORo project we denote by exported data all the data that is generated from the core data of the project.

Usually, data is exported from ROCC each time a critical mass is reached (see Section 4.2), in order to perform training and evaluation of deep-learning models. The pipeline for export of core data in the format specified by the export schema is under development.

4 *Collecting Data*

The previous section detailed the objects generated by the OOCIAT interface¹⁴ and/or the machine, as a result of page image manual annotation or automatic segmentation and recognition operations. The difference between manually marking an object on a page (done by an annotator) and the one performed automatically is recorded in the @objectAnnotator field, which can be the ID of a human annotator or AUTOMATIC. The existence of these two types of values makes the distinction between the segmentations done manually (gold), and those done automatically (test), and will make it possible to evaluate, by comparison, the technology of object identification (OI) and character recognition (CR), for each object type.

4.1 *Manual Data Acquisition*

Manual annotation refers to two components: visually identified objects, i.e., columns, lines, words, characters, and their Latin transcriptions. The dimension and position per page of the objects (framed in rectangles), as well as their content (the Latin transcription) are being noted by humans in the XML notation, in editing sessions, facilitated by the OOCIAT interface. To accelerate the

⁹ We consider the text to be obstructed when at least one character cannot be read.

¹⁰ As, for instance, when red is shifted from black, with partial coverage.

¹¹ The same sheet of paper was reused for writing a new text on it after erasing old text but with traces of old text remaining on the page.

¹² We classify the corrections on the page according to the following scale: more than 5, between 1 and 5, or none.

¹³ Idem - with more than 5 marginal notes, between 1 and 5, or none.

¹⁴ The *Online Old Cyrillic Image Annotation Tool*, see [Pădurariu, 2021].

acquisition process, the DeLORo consortium decided that the identification of a limited set of Cyrillic characters can also be carried out by personnel with a different background than linguistics, without the danger of introducing errors¹⁵. There are several letters with multiple values and those are annotated only by linguists. The characters which have multiple values are: Ъ (ă, u, î), Ь (ă, u), ъ (î, u), А (ia, ea); also, in certain contexts, the characters (Ѣ, ѣ) may be non-phonetic, therefore have void transcriptions. The output of acquisition sessions with this interface will be used to acquire the gold corpus.

Creation of metadata for the corpus content is also a very important activity for this project. This activity provides details about: language (mainly Romanian, but short fragments in other languages could accidentally occur), year of document publication, document printing house (for prints), the calligrapher (for uncials, if mentioned), author, translator, free description of noise (degraded pages, ink stains, creases, dirt, etc.). In addition, details such as supra- and/or under- linear writing, the existence of any critical edition of the text (with indication of source), the level of processing complexity required, the type of document script, the level of annotation, etc. are annotated by the members of the team. Coding of metadata and its acquisition was an important activity all along building the CoRoLa corpus [Cristea *et al.*, 2017] and we have exploited here the high expertise for this type of activity of some of the members of the Institute of Computer Science from ARFI.

4.2 Shortcutting Data Acquisition

In [Cristea *et al.*, 2000], an algorithm aiming to boost manual annotation of primary documents, is briefly sketched. Such an approach is motivated by the difficulty of acquiring enough data during the project lifetime that would safely support the training of the CR technology, supposing the lighter technology of OI works satisfactory¹⁶. Instead, we exploit a set of documents already transcribed by experts, trying thus to shortcut the data acquisition process.

In this section we detail this process. First, we introduce some notations. Let *IMAGE* be a scanned page (characterised by a page ID in the ROCC space) of a Cyrillic Romanian document, for which the database includes a Latin transcription. It contains a set of rectangular shaped frames of objects of type character (as *core data*), with sides parallel to the horizontal and vertical axis, each object being characterized by: a unique ID, the coordinates (horizontal and vertical) of two points (left-top, right-bottom), and a label (which can be VOID, meaning unrecognised, or a Latin letter). The frames of these objects (called

¹⁵ The team of consortium members that works with the OOCIAT interface includes 4 linguist experts and 3 non-linguists. The non-linguists have been attentively trained by experts to recognise and transcribe isolated Cyrillic characters.

¹⁶ We do not leave aside also the alternative of recognising Cyrillic characters directly from the line frames, without identifying character frames.

“character containers”) have been automatically identified by the OI technology, while their contents are the result of applying the CR technology.

Let TEXT be a one-page ASCII coded text, representing the interpretative transcription of the IMAGE (as *imported data*); the characters in TEXT are ascendingly sorted by their offsets. Texts of pages are delimited by special <p> markings placed in-between them.

As such, the *alignment space* is a set of pairs IMAGE-TEXT.

The shortcutting strategy

This strategy is applied iteratively to perform cycles of annotation-training-evaluation-alignment, with the goal of gaining more training data than directly annotated by humans, by exploiting the texts manually transcribed by experts. It can be applied at different stages along the painful and tedious process of manual data annotation (through the specially designed OOCIAT interface). Each training is followed by an assessment of the accuracy of recognition of all objects of type <object:Character>.

From the experience reported so far [Ren *et al.*, 2015], a scale of ascending numbers of instances required in a training process can be established. This is because, for different types of objects, the quantity of training instances leading to the same recognition accuracy score can be different (for example, the <object:Character> type can have the scale: 1,000, 5,000, 10,000, 20,000, etc.). Then, a threshold for the recognition accuracy of a character is set, let's say θ (e.g., 97%), above which the recognized characters are considered to be “safe”.

Initialisation:

1. Build the LatinChars data structure, as a set of objects, of cardinal equal to the set of Latin characters, each instance having the structure: {char C, list imCSet, integer count, float acc, integer nextLevel}, where: C holds a Latin character value, imCSet is the index of all image instances representing the character C in the training set (stored as IDs of objects of type <object:Character>), count is the cardinal of imCSet, i.e. the training set for recognizing C, safeAnnoLevel represents the next value on the scale of quantities of training instances that should be reached to acquire the θ threshold, and acc will be updated with the accuracy of recognition of the character C, as measured after each training iteration (in the range: 0 - 100%); initially, leave these values empty.

2. Propose and then calibrate a scoring function able to appreciate the confidence in aligning a container from IMAGE with a TEXT character and an algorithm that proposes a set of ALIGNMENTS of the form <c, container-ID, off-char, s>, where: c is a character value, container-ID is the ID of a container in IMAGE whose value is c (different from VOID), off-char is the offset of a character c in TEXT, and s is the value of the scoring function for this alignment.

3. To perform the training and evaluation, divide the *core data* of type <object:Character> in ROCC in two disjunct sets: *gold*, *test*, by applying a percentage distribution (for instance 90%, 10%).

4. Initially train the CR neural network on the *gold* data, then evaluate and set the acc values of the LatinChars structure to express the average accuracy of each Latin character.

As will be seen, the CR neural network will be periodically trained on more and more instances of the set LatinChars, namely each time the number of gold objects of a certain type, accumulated by hand annotation and/or resulting by applying the shortcutting strategy, exceeds the next safeAnnoLevel for that character. Therefore, the level of accuracy acc, reached for each element in the LatinChars set, is updated after each training + evaluation session.

An occurrence in IMAGE of an instance of the <object:Character> class can be classified in two ways: as a character container with unknown content, let's call it <CharacterVoid>, if the neural network reports a confidence of recognition for that instance lower than a threshold (in this case the @objectContent attribute is set to VOID), or as a container carrying a labelled content, let's call this instance <CharacterX>, where X is a Latin letter encoding the character (in this case the @objectContent attribute is set to X).

5. For each IMAGE-TEXT pair in the *alignment space* do the following:

Linearization:

6. Sort the containers found in IMAGE left-to-right, by using the centres of gravity of their shapes, approximately aligned on horizontal lines, and then concatenate the consecutive lines to produce one single string of character Ids.

7. Output the string obtained by transcribing in order the labels of containers and replacing the VOID label by '\$'. When the horizontal distance between two containers is greater than a threshold determined experimentally, propose the generation of a space, with a certain degree of confidence (score).

Alignment:

8. Search IMAGE for a pair of containers, <CharacterX> and <CharacterY>, both with $\text{acc} > \theta$, as reported by the corresponding LatinChars elements, between which there are d (small, e.g., between 1 and 10) other intermediate containers of type <CharacterVoid>, whose $\text{acc} \leq \theta$ (spaces are not counted). This should produce one 10-uple $\langle pI, X, imX, offIX, offTX, pT, Y, imY, offIY, offTY \rangle$, where pI is the Id of an IMAGE page where the sequence was discovered, pT - that of the corresponding TEXT page, X and Y are the two characters, imX and imY are the Ids of the containers <CharacterX> and <CharacterY> in the IMAGE page pI , $offIX$ and $offIY$ are the offsets of imX and imY in the same IMAGE page, $offTX$ and $offTY$ are the offsets of the two characters in the TEXT page pT , $|offY - offX| = d + 1$, in-between the two characters X and Y in the TEXT page there exist the same number d of interposed letters, and $offIX$ and $offIY$ are approximately equal with $offTX$, respectively $offTY$.

9. Align all characters in IMAGE placed between imX and imY , one to one, with the corresponding TEXT characters placed between X and Y , which means:

- each image-character alignment is a pair $\langle imC, C \rangle$;

- for each aligned character C in the LatinChars structure increment the count, and add imC to the imCSet setlist;

10. Check if in the LatinChars set new characters whose counters count exceed safeAnnoLevel appear (this is the next threshold of number of instances on the scale of the training levels, as described at step 1, that should be reached in order to start a new training process). If yes:

- resume training for pairs <imC, C> (with their training sets imCSet) whose counts exceed their safeAnnoLevel thresholds;
- for each newly trained C character, update:
- acc to the value resulting from the evaluation of the accuracy on the new gold sets;
- safeAnnoLevel to the next threshold in the list at step 1.

11. If all characters in the LatinChars set have reached their maximum training thresholds (safeAnnoLevel), then exit.

12. If no new "safe" characters have appeared in the LatinChars set, which now includes new accuracy values, announce the message "More data needed!".

13. Otherwise, go to step 5.

5 Accessing Data

5.1 Techniques for Searching and Focusing Information in Historical Documents

Using as input the images and their annotations made by specialists (frame coordinates and textual values) and a sequence of words whose occurrence the user would like to identify inside the images, we aim to create an algorithm that can identify and focus text in the image. Knowing the rectangular frame a line of text is positioned into and the frame bordering each character, first the characters belonging to a line can be detected and ordered left-to-right, thus extracting the text belonging to each line. Then, after concatenating the strings belonging to each line, the resulting text could be queried by searching inside it for the words given as input. When the sequence of words belonging to the query is identified within the transcribed text, using back the links between the page displaying the image of the original Cyrillic document and the transcribed Latin text, a surrounding area in the image can be computed and that zone displayed as the result of the query, on which a highlight could put in evidence the exact occurrence of the searched string of words.

This method may not always lead to a relevant result, as differences may appear between the concatenated string of characters of the query and the interrogated strings in the transcribed text. Differences may be caused by: errors in the CR flow

or different morphological values of flexing words (verbs, nouns, adjectives, pronouns). At this moment we believe that an approximate string-matching algorithm intermediating the search, which would make use of the editing distance, might represent a good compromise for a solution. In our experiments we considered a window, equal in length or smoothly longer than the length of a word to be found. At each intermediate position of this window, moving from left to right over the interrogated string, the editing distance between a query word and the characters inside the window is computed. If this distance is greater than a threshold, experimentally established, then the word can be considered correctly identified and the search should move on to the next word of the querying string.

With this solution in place, more complex types of searches can be thought of, for instance combining string search with constraints over the searched collection of documents, which are rooted in metadata (author, title, time period, etc.).

5.2 *ROCC Statistics, as per Today*

The lexicon is of tremendous importance in our enterprise, because, through language-based models, it can influence the recognition process. Moreover, following the Cyrillic Romanian transliteration stages, the word should be placed at the bases of any process of user enquiring over the old documents. This is why, in the second part of this section we will refer to the lexicon acquired till now from ROCC. But we will see first an initial round of statistics (because data acquisition is in a continuous process) of the annotated data acquired till now.

Table 1 Number of annotated instances per object type

Object type	Annotated instances	Object type	Annotated instances	Object type	Annotated instances
Accolade	0	Column	7	Ornate letter	374
Frontispiece	30	Letters	146.803	Lines	6.923
Marginal writing	47	Modifier	5.057	Ornament	82
Out-of-line letter	1.309	Ref. above line	0	Title	157

In the construction of the lexicon of the ROCC corpus, two data sources are used, combined: (i) the annotations of lines operated in OOCIAT, and (ii) the texts

of those documents for which there exist complete transcriptions. Because, according to the definition of the corpus schema, the line annotations are stored in the database and the text of the transcripts - in files on disk, the data extraction process is divided into two steps. In the first step, the contents of OOCIAT-annotated lines are extracted from the database, together with the year of publication of the work the lines belong to. In the second step, for each transcribed work, the year of its publication is extracted from the database, together with the disk address of the file containing the transcribed text. Then, strings of characters acquired from both sources (i) and (ii) are concatenated, after filtering out documents for which incomplete metadata do not indicate a year¹⁷. A model of the Spacy library¹⁸, trained on the modern Romanian language¹⁹, was used to tokenize this file and to eliminate punctuation marks. Each of the terms identified by the Spacy model was subjected to an additional validation process that removed tokens containing numbers and those split by a hyphen at the end of a row. The resulting list of word occurrences is then split into more sub-lists, each containing terms in a period of 50 years. The statistics are resumed in Table 1 and Table 2 (no data is available for the period 1600-1700).

Out of the data collected until the time of writing this paper, we could identify 918 terms that appear in at least two periods. The process of obtaining the vocabularies by period and computing statistics on them is automated in a Python script²⁰.

There are several directions to continue this activity. First, we will run the algorithm described on all the documents included in the ROCC corpus, in order to get a more comprehensive idea of the old Romanian lexicon included in our imported data. Then, we intend to also use the existing metadata stating the place of appearance of the documents, preparing this way for synchronicity studies. Finally, we intend to combine the lexicon thus prepared with others from various sources, such as, for example, the one of *Monumenta lingua Dacoromanorum - Biblia 1688*²¹ and the one of *Manuscript 45* [Munteanu *et al.*, 2016].

¹⁷ Metadata for a great part of the original documents, in XML format, was received from the Library of the Romanian Academy in Bucharest.

¹⁸ <https://spacy.io>

¹⁹ https://spacy.io/models/ro#ro_core_news_lg

²⁰ The source code is available at <https://github.com/deloro-project/rocc-pipelines>. Although the execution of this script can be triggered via Github Actions without access to the server containing the database, the results can only be downloaded with an active user account: <https://github.com/deloro-project/rocc-pipelines/actions/workflows/export-lexicon.yml>

²¹ Iași, tom. I-XXIV, 1988-2014. Coordinators of the project: Vasile Arvinte, Alexandru Andriescu, Paul Miron, Eugen Munteanu.

Table 2 No. of terms in the lexicon for each 50-years period

Period	No. terms	Period	No. terms
1550-1600	2021	1750-1800	802
1700-1750	4099	1800-1850	1244

6 Conclusions

In this paper we described the organisation of a database supporting the DeLORo project and its accessing facilities. Nothing has been said about the DL technology that has to be used for OI and CR, key components of the whole enterprise. These must clearly make the subject of another paper. Known as very data hungry, we considered of primary importance the organisation of data and the way these data that will have to be used in training, calibration of networks and evaluations should be acquired.

By its complexity, our approach only partially resembles an OCR-like one, because we collect data from the image at a very high level of detail. The choice for this rather sophisticated organisation of data is induced by a number of challenges, by which we intend to copy on the machine the exquisite abilities of linguistic experts to decipher Cyrillic Romanian texts. We mention here a few: the need to interpret in context the meaning of some multiple valued Cyrillic letters (interpretative transcription); the issue of recognising and inserting at their places characters appearing outside the lines, as those written in-between them or on the margin of the document; applying transformations on the values of characters, as indicated by modifiers; putting in correspondence reference marks on margins with those placed above lines; segmentation of words when separating blanks are missing; the need to focus segments of Cyrillic texts in the original scans of pages in a reverse Latin-to-Cyrillic user-commanded search. Departing from the basic character details, up to more elaborate layers of processing the old Romanian language, the project also addresses the issues of word segmentation and lemmatisation, and will try to combine recent studies of syntax and semantics to the data acquired through automatic transcription. All these issues have solutions rooted in the data organised as described in this paper, but their actual elaboration should be detailed in the forthcoming activities of the DeLORo project.

Acknowledgements

This work has been carried out with the financial support of the Romanian Executive Unit for Financing Higher Education, Research, Development and

Innovation (UEFISCDI), within the project *Artificial Intelligence Models (Deep Learning) Applied in the Analysis of Old Romanian Language (DeLORo – Deep Learning for Old Romanian)*, project code PN-III-P2-2.1-PED-2019-3952.

We thank our non-author colleagues in the project: Cecilia Speranța Bolea, Horia Radu Borza, Paula Alexandra Crucianu, Mihaela Găman, Daniela Gîfu, Gabriela Haja, Radu Ionescu, Marius Popescu, Elena Tamba, Roxana Vieru, engaged in data acquisition and modelling DL technologies.

We thank the curators from the Library of the Romanian Academy in Bucharest, Gabriela Dumitrescu and Marilena Bănică for providing most of the primary documents and their XML metadata.

References

- Bianu I., Hodoș N., Simionescu D. (1903-1944). Bibliografia românească veche. 1508–1830, Vol. I–V, Edițiunea Academiei Române, București, 2490 p.
- Cândea V. (2011, 2012, 2014). Mărturii românești peste hotare, Editura Biblioteca Bucureștilor, Editura Academiei Române, Editura Muzeului Literaturii.
- Cristea D., Simionescu R., Haja G. (2012). Reconstructing the Diachronic Morphology of Romanian from Dictionary Citations. In Proceedings of LREC-2012, Istanbul, 21-25 May.
- Cristea D., Gîfu D., Moruz A., Onofrei M., Pistol L., Scutelnicu L., Bolea C. (2017). An Insight into the Corpus of Contemporary Romanian. In *Memoirs of the Scientific Sections/Memoriile Secțiilor Științifice*, Series IV, Tome XL, Publishing House of the Romanian Academy, ISSN 1224-1407, pp. 67-84.
- Cristea, D., Gîfu, D., Cojocaru, S., Colesnicov, A., Malahov, L., Popescu, M, Onofrei. M., Bolea. D.C. (2018). How To Find Out Whether the Romanian Language Was Influenced by the Two Historical Unions?, in V.Păiș, D.Gîfu, D.Trandabăț, D.Cristea, D.Tușiș (eds.) *Proceedings of the 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language*, noiembrie 22-23, Editura Universității “A.I.Cuza” din Iași, pp. 77-88, ISSN 1843-911X.
- Cristea, D., Pistol, I., Gîfu, D. (2018a). Connecting the Content of Books to the Web and the Real World. In *Proceedings of the 19th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing 2018, 18-24 March 2018, Hanoi, Vietnam.
- Cristea, D., Pădurariu, C., Rebeja, P., Onofrei, M. (2020). From Scan to Text. Methodology, Solutions and Perspectives of Deciphering Old Cyrillic Romanian Documents into the Latin Script. In *Knowledge, Language, Models, Volume in Honour of Prof. Galia Angelova on Her 65th Birthday*. Milena Slavcheva (ed.), INCOMA Ltd. Shoumen, Bulgaria. ISBN 978-954-452-062-5, pp. 38-56.
- Dumitrescu, G. (coord.) (2017). Biblioteca Academiei Române 1867-2017. Tezaur de cunoaștere. 150 de ani, București.
- Dumitrescu G. (2018). Canon de pocăință cuprinzând povestea sfinților osândiți ai Scării, Editura Excelența prin Cultură, București, pp. 128.
- Gafton A. (2016). Reflecții asupra principiilor de transcriere a vechilor texte românești, In *Studii de filologie. In Honorem Ștefan Găitânaru*, Editura Universității din Pitești,

- 2016, pp. 185-194, online at: https://www.academia.edu/28670753/Reflec%C8%9Bii_asupra_principiilor_de_transcriere
- Gîfu D. (2015). Contrastive Diachronic Study on Romanian Language. In *Proceedings FOI-2015*, S. Cojocaru, C. Găindric (eds.), Institute of Mathematics and Computer Science, Academy of Sciences of Moldova, pp. 296-310.
- Gîfu D. (2016). Lexical Semantics in Text Processing. Contrastive Diachronic Studies on Romanian Language, In *PhD Thesis*, “Alexandru Ioan Cuza” University of Iași, Romania.
- Gheție I., Mareș A. (1985). Originile scrisului în limba română. Editura Științifică și Enciclopedică, București, pp. 463.
- Ionescu R.T., Popescu M., and Cahill A. (2016). String kernels for native language identification: Insights from behind the curtains. In *Computational Linguistics*, 42(3), pp. 491-525.
- Munteanu, E. (coord.), Gînsac, A.-M., Minuț, A.-M., Munteanu, L.-G., Ungureanu, M. (2016). Vechiul Testament Septuaginta. Versiunea lui Nicolae Spătarul Milescu (Ms. 45 de la Biblioteca Filialei Cluj a Academiei Române), Editura Universității “Alexandru Ioan Cuza” din Iași.
- Nastase, V., Popescu, M. (2009). What's in a name? In some languages, grammatical gender. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol. 3, pp. 1368-1377.
- Popescu, M., Hristea, F. (2011). State of the art versus classical clustering for unsupervised word sense disambiguation. *Artificial Intelligence Review*, 35(3), pp. 241-264.
- Plămadă-Onofrei M., (2018). De obște gheografie, de Amfilohie Hotiniu, Iași, 1795. Ediție de text, studiu filologic și lingvistic, In *PhD Thesis*, “Alexandru Ioan Cuza” University of Iași, Romania.
- Ren, S., He, K., Girshick, R. and Sun, J. (2015). “Faster R-CNN: Towards real-time object detection with region proposal networks”. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, Volume 1, ser. NIPS'15, pp. 111-121.
- Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks, in *NIPS: Neural Information Processing*, Volume 2, December, pp. 3104–3112.

Preservation of Romanian Linguistic Heritage. Framework for Dialectal Data Management

SILVIU-IOAN BEJINARIU¹, FLORIN IFTENE^{1,2},
MANUELA NEVACI¹, and CARMEN IRINA FLOAREA¹

¹ *“Iorgu Iordan – Al. Rosetti” Institute of Linguistics, Romanian Academy*

² *Institute of Computer Science, Romanian Academy Iași Branch*

{silviu.bejinariu,
florin.iftene}@iit.academiaromana-is.ro,
{elanevac, flo_irina}@yahoo.com

Abstract

Dialectology studies the variation of language especially in terms of geographical distribution, but also its evolution over time. Romanian linguistic cartography has a tradition of over a century being one of the oldest and richest in Europe. In the last two decades, the dialectal data began to be stored in electronic format, in databases which allows their use for other research or analysis in addition to their publication in linguistic atlases. In this paper a framework for dialectal data management is proposed. In addition to the usual information, i.e., notions, survey points and answers received during the interviews, the relational database also allows the storage of additional information, mainly multimedia. Also, it allows storing general (not personal) information about the interviewee and answers provided in the same survey point by several people, at the same or different times. The proposed model is used for the elaboration of the Linguistic Atlas of Romanian Dialects from the North and South of the Danube and creates the premises to store in a single database the dialectal information from all the Romanian linguistic atlases.

1 Introduction

Romanian linguistic cartography has a tradition of over a century being one of the oldest and richest in Europe. In the early days, the answers from dialectal surveys were noted on paper by the investigators and then transcribed in the published works. Later, the answers began to be audio recorded and then interpreted and transcribed. Most of the dialectal resources from before 2000 are still kept only in these formats handwritten, printed or as audio recordings, which makes it difficult to reuse them in other studies. In the last two decades, research tools based on the use of information technologies have begun to be used in dialectology. The newly collected dialectal data are stored in electronic format, in databases which allows their use for other research or analysis in addition to their publication in linguistic atlases.

The framework for dialectal data management proposed in this paper is a development of previous applications and data structures that have already been used in the elaboration of some linguistic atlases described in the next section. In the proposed model, the usual information, i.e., notions, survey points and answers received during the interviews stored in the relational database is completed with the storage of multimedia files: audio / video recordings, images and text related to the interview, to the localities where the surveys take place, or to other elements which are relevant for the dialectal studies. The general information about the speakers and the date of the investigation allows answers provided in the same survey point by several people, at the same or different times to be stored in the database. Thus, the premises are created to store in a single database the dialectal information from all the Romanian linguistic atlases.

All over the world there is a long tradition related to the study of geolinguistics which materializes through the publication of dialectal material in the form of phonetic or interpretive maps in linguistic atlases. Among the most remarkable achievements in the European space, the following must be mentioned: *Atlante Linguistico Italiano*¹ (ALI), *Atlas Lingüístico de la Península Ibérica*² (ALPI), *Linguistic Atlas of the Romance Domain* (ALiR), *General Slavic Linguistic Atlas*³ (OLA).

*L'Atles Lingüístic del Domini Català*⁴ (ALDC) is a dialectology project developed at the Institute of Catalan Studies from Barcelona and it is dedicated to the Catalan language. It is published in three printed formats: *The Linguistic Atlas of the Catalan Domain*, *The Small Linguistic Atlas of the Catalan Domain* and *Ethnotexts*. Another important achievement in the domain is the *Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects*⁵ (ALD). It consists of two parts which were published in 14 volumes with about 2950 geolinguistic maps. It is accompanied by a DVD which contains the world's first "Speaking atlas" with 21 investigation points and 806 questions (ALD). An extensive work is also *Atlas Linguarum Europae*⁶ (in English: *Linguistic Atlas of Europe*) (ALE) which is elaborated by dialectologists from more than 47 countries.

Sousa [2017] presents the structure and design of the database used by the *Atlas Lingüístico Galego* to store the materials. The database is not only the source of information for the atlas publication but also for other research related to the variation and change of Galician varieties, and other projects with a larger scope. Hessle and Kirk [2020] present the advantages of a digital atlas, such as the

¹ <https://www.atlantelinguistico.it/atlante/introduzione.html>

² <http://alpi.csic.es/en>

³ <http://ola.zrc-sazu.si/index.htm>

⁴ <https://aldc.espais.iec.cat/>

⁵ <https://www.ald.gwi.uni-muenchen.de/en/?db=ald2>

⁶ <https://lingv.ro/atlas-linguarum-europae/>

possibility to visually contextualize the linguistic data by combining it with extra linguistic information in the *Linguistic Atlas of Scotland*. The optical character recognition is also used for automatically digitizing the old linguistic materials. Another area of research is the digitization of old linguistic atlases. Scherrer and Kellerhals [2014] present the results obtained in digitization of the linguistic atlas of German-speaking Switzerland that allows quantitative analysis and computational processing of linguistic data. The work was continued by the creation of the *ArchiMob*⁷ corpus of linguistic varieties spoken within the territory of Switzerland. *ArchiMob* contains samples of transcribed text in Swiss German, intended for quantitative approaches to the study of language use and variation [Scherrer *et al.*, 2019].

The paper is organized as follows. In the second section are briefly described the main works of linguistic geography published in Romania. The most important achievements related to the computer aided design of the linguistic maps are described in the third section. In the fourth section are described the main improvement proposed for the data management component in order to publish the *Linguistic Atlas of Romanian Dialects from the North and South of the Danube* (in Romanian: *Atlasul lingvistic al dialectelor românești din nordul și din sudul Dunării*, abrev. *ALDRo*). The conclusions are presented in the last section of the paper.

2 *Linguistic Atlases in Romania*

Although the first studies of Romanian dialectology are much older, the first linguistic atlases began to be published only a century ago. Over time, the technology of making linguistic maps has evolved from maps published using offset technologies to those made using the information technology methods, published both in printed format and online on the Internet [Bejinariu *et al.*, 2021].

The first linguistic atlas was published by the German romanist Gustav Weigand between 1898 and 1909 and it has nine issues: *Linguistischer Atlas des dacorumänischen Sprachgebietes*⁸ (WLAD). This was the first systematic approach in the study of the Daco-Romanian dialect. After 1919, was started the elaboration of the *Romanian Linguistic Atlas* (in Romanian: *Atlasul Lingvistic Român*, abrev. ALR) at the initiative of Sextil Pușcariu. ALR was conceived as consisting of two parts: *ALR I* and *ALR II*, these having complementary networks of points and different investigators: Sever Pop and Emil Petrovici [Olariu and Olariu, 2010]. Among the investigation points are also some that correspond to the Aromanian, Megleno-Romanian and Istro-Romanian dialects. Even if the

⁷ <https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html>

⁸ https://www.europeana.eu/en/item/9200456/handle_123456789_54079

investigations were done, due to World War II, only three of the ten proposed volumes appeared: two volumes of ALR I and only one of ALR II. Two supplements were published for ALR II, of which one contained dialectal texts in phonetic transcription. Also, each of the three ALR volumes with analytical maps (Fig. 1) were accompanied by a volume called the *Small Romanian Linguistic Atlas* (in Romanian: Micul Atlas Lingvistic Român, abrev. ALRM) which contains color synthetic maps (Fig. 2).



Figure 1. Map 228, ALR I Vol. I (source: Digital Library of the Institute of Linguistics and Literary History "Sextil Pușcariu", Cluj-Napoca⁹)

From the point of view of graphic design, it is noted that the template established in ALR for the linguistic maps has been preserved and is still used in the new atlases with very few changes. Additional information is organized in three *Notes* containing the question for which the title word represents an answer, (*Note I*), other information / comments obtained from the speakers (*Note II*) and other explanations / comments of the author or investigator (*Note III*), as in the actual linguistic atlases.

⁹ <http://inst-puscariu.ro/alr.html>

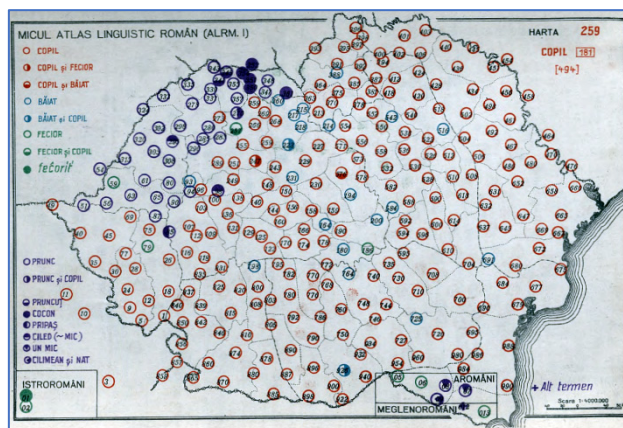


Figure 2. Synthetic map 259, *ALRM I* vol. II, (source: Digital Library of the Institute of Linguistics and Literary History "Sextil Pușcariu", Cluj-Napoca¹⁰)

After the second world war, the unpublished material from *ALR I* and *ALR II* was capitalized under the leadership of Emil Petrovici by the Institute of Linguistics from Cluj-Napoca. Thus, seven volumes of the *Romanian Linguistic Atlas – new series* (in Romanian: *Atlasul Lingvistic Român. Serie nouă*, abrev. *ALR-SN*) were published between 1956 and 1972 and four volumes of the *Small Romanian Linguistic Atlas - new series* (in Romanian: *Micul Atlas Lingvistic Român. Serie nouă*, abrev. *ALRM-SN*) were published until 1981.

In 1957 Emil Petrovici proposes the publication of the *New Romanian Linguistic Atlas by Regions* (in Romanian: *Noul Atlas Lingvistic Român pe Regiuni*, abrev. *NALRR*). For a more detailed analysis, the investigation network was enriched to about 1000 survey points, but the number of questions in the questionnaire (~2600) was reduced comparing to *ALR*. The investigations were carried out by the dialectologists from Bucharest, Cluj-Napoca, Iasi and Timisoara between 1960 and 1980. *NALRR* was designed to contain seven regional atlases for the Daco-Romanian dialect which corresponds to the historical regions: *Muntenia și Dobrogea*, *Oltenia*, *Banat*, *Crișana*, *Maramureș*, *Transylvania* and *Moldova and Bukovina*. Since 1967, when their publication started, four or five volumes were published for each of the seven regions and it continues. Being different groups of authors, each of them tried to bring various improvements in the way of presenting and systematizing the linguistic material, most of them being considered in the volumes published later: synthetic and phonetic maps, morphological or lexical interpretations, traditional linguistic maps with phonetic

¹⁰ <http://inst-puscariu.ro/alr.html>

transcripts combined with isoglosses, different ways of grouping and ordering the answers in plates with (un) mapped material [Saramandu, 2007]. The above-mentioned works are not the only linguistic atlases of the Romanian language. The *Moldavian Linguistic Atlas* (in Romanian: *Atlasul Lingvistic Moldovenesc*, abrev. ALM) was published between 1968 and 1987 using the Cyrillic alphabet. After 1990, with a new title, *The Romanian Linguistic Atlas by regions. Bessarabia, Northern Bukovina and Transnistria* it was edited using the Latin alphabet and “it is part of the series of these regional atlases” [Pavel, 1993].

Other works dedicated to the study of the dialects of the Romanian Language are: *The linguistic atlas of the Romanian dialects from the Yugoslav Banat* by Radu Flora, the *Small linguistic atlas of the Istro-Romanian dialects* by Radu Flora, the *Small Atlas of the Aromanian dialect from Albania and the Former Yugoslav Republic of Macedonia* by Petre Neiescu, *The Linguistic Atlas of the Megleno-Romanian Dialect* by Petar Atanasov and *The Romanian linguistic atlas. The dialects between Morava, Danube and Timoc* published by the Institute of Linguistics and Literary History “Sextil Pușcariu” in Cluj-Napoca [Saramandu and Nevaci, 2013]. *The Linguistic Atlas of the Aromanian Dialect*¹¹ (ALAR) by Nicolae Saramandu and Manuela Nevaci (editor) is also a work of which two volumes have already been published. It is based on the NALR questionnaire with 2543 questions and 61 investigation points from Romania, Bulgaria, Republic of North Macedonia, Greece and Albania [Saramandu and Nevaci, 2018].

Saramandu [1990] posed “the problem of elaborating an atlas of regional linguistic atlases or a synthesis of the New Romanian linguistic atlas by regions”. Two volumes with interpretative maps were published under the title *Romanian linguistic atlas by regions - Synthesis* (in Romanian: *Atlasul Lingvistic Român pe Regiuni. Sinteza*, abrev: ALRR. Sinteza) coordinated by Nicolae Saramandu. The network consists of about 1200 investigation points, including those from the Republic of Moldova (ALRR.Sinteza).

Most of the works mentioned above were prepared using traditional techniques, respectively the content was drawn by specialized cartographers. The use of IT solutions for the elaboration of linguistic atlases started after 2000.

3 Computer Aided Design of the Linguistic Maps

The linguistic atlases published in the last two decades were made using computerized methods. There were two types of approaches. In the hybrid approaches the linguistic material digitized in Word documents using specialized fonts and then the captured images of phonetic transcriptions are placed on the map template, e.g., the first volume of *ALAR*. The second approach is fully

¹¹ <https://lingv.ro/wp-content/uploads/2021/11/Atlasul-lingvistic-al-dialectului-aroman-ALDA.pdf>

computerized: the linguistic material is digitized and encoded in specialized databases. The maps are generated automatically and they can be edited later only to improve the aesthetic aspect. This second approach was used for the publication of *NALRR. Crișana* volumes III and IV; *NALRR Moldova and Bukovina* volumes III and IV; and *ALAR* vol. II. In the last period, to recover and make it available, the Institute of Linguistics and Literary History „Sextil Pușcariu” from Cluj-Napoca started the operation of scanning and publishing on the internet of the old atlases published using traditional methods.

Between the fully computerized approaches there are two methods of making maps: (a) the maps are generated automatically using specialized applications developed for this purpose and (b) Geographic Information Systems packages are used. The first method is required in case of phonetic maps because there are no fonts to include all the possible symbols used by the traditional phonetic transcription of the Romanian language while the second is used in case of interpretative maps, when the answers are indicated using graphical symbols.

3.1 *Specialized Applications for Linguistic Maps Generation and Editing*

Specialized applications for linguistic atlases preparation began to be used after 2000 and usually they are specific to a certain atlas. The general structure is presented in Fig. 3 and two main databases are used. The geographical database contains the map template and the fixed components of the maps and it is prepared using a GIS package. The specialized application was developed for the linguistic database management including the phonetic transcriptions editing, automatic generation and editing of the linguistic maps and (un) cartographed materials.

The first system for computer based preparation of linguistic atlases in Romania is the *ALR* application which used for publishing volumes III and IV of *NALRR. Moldova and Bucovina* (abbrev. NALRR. MB). It was developed in the early 2000s and it is the result of the collaboration between researchers from the Institute of Computer Science (ICS) and the „A. Philippide” Institute of Romanian Philology (IFR) of the Romanian Academy Iasi Branch. *ALR* application allows editing phonetic transcriptions with all possible combinations of phonetic phenomena using only two TrueType fonts and an image synthesizer [Apopei *et al.*, 2002]. The linguistic database allows the storage of an audio recording of the answer received during the investigation.

The second version of the *ALR* application, named *AlrMaps* was developed by ICS in collaboration with „Iorgu Iordan – Al. Rosetti” Institute of Linguistics of the Romanian Academy (ILIR) for the preparation of the second volume of the *Linguistic Atlas of the Romanian Dialect*¹² (*ALAR*) [Bejinariu *et al.*, 2018]. The

¹² <https://lingv.ro/wp-content/uploads/2021/11/Atlasul-lingvistic-al-dialectului-aroman-ALDA.pdf>

main improvement is that *AlrMaps* can be used to create any linguistic atlas specific to the Romanian space by modifying certain application settings. Both versions include also specialized editors for dialectal texts which are published in socio-linguistic atlases.

Another approach is the *Romanian Online Dialect Atlas* application which was developed (with two versions, RODA I and RODA II) for the publication of volumes III and IV of *NALRR. Crișana* [Embleton *et al.*, 2013]. It is the result of the collaboration between the research group led by Dorin Uritescu from York University, Toronto and the Institute of Linguistics and Literary History "Sextil Pușcariu" from Cluj-Napoca. One of the novelties brought by the RODA systems is the creation of dialectometric interpretive maps.

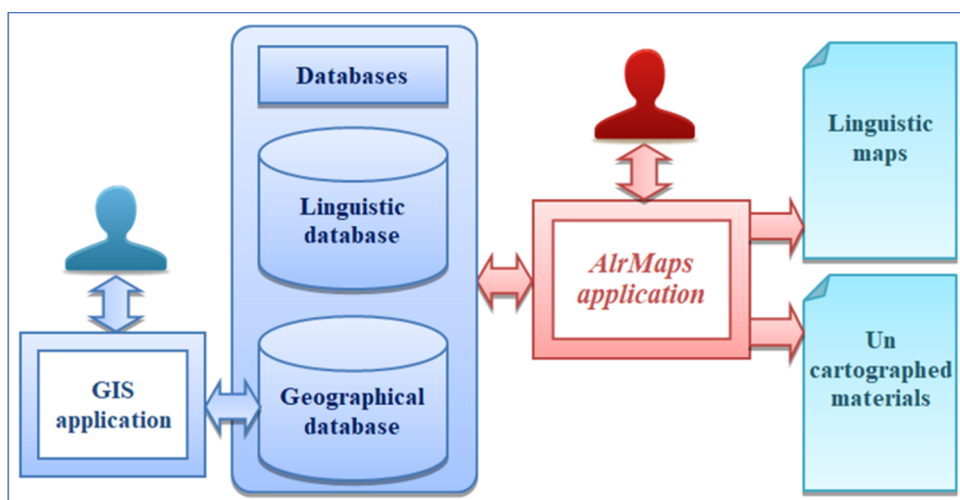


Figure 3. General structure of the specialized packages for linguistic maps preparation.
(source: [Bejinariu and Olariu, 2017])

A special achievement in linguistic cartography, published only in the web version, is the *Audiovisual Linguistic Atlas of Bukovina*¹³ (in Romanian: Atlasul lingvistic audiovizual al Bucovinei, abrev. ALAB). This is the first audio-video atlas centered on sociolinguistic features [Olariu and Olariu, 2014]. Unlike other linguistic atlases, ALAB is published only on internet and it does not present a phonetic or interpretative map. On the project's site is displayed only the image of the map with the investigation points colored according to the variation (age and sex) of the selected speaker. The associated transcriptions are displayed

¹³ <http://www.philippide.ro/alab/rezumat.html>

simultaneously with the video sequence showing the speaker's response after selecting the survey point on the map [Olariu *et al.*, 2016].

3.2 GIS Based Systems for Linguistic Maps Preparation

The GIS based applications are dedicated to linguistic atlases which contain only interpretative maps in which the language variation is shown using graphic symbols. As it is presented in Fig. 4, it consists of two main components: (a) the legend (database) management tool and (b) the map generator. Even if the two components are implemented as a single application, the functionality was divided to allow the maps generation using the old legends (in Word or Excel files) used in previous versions of the atlases without being necessary to digitalize again the linguistic material.

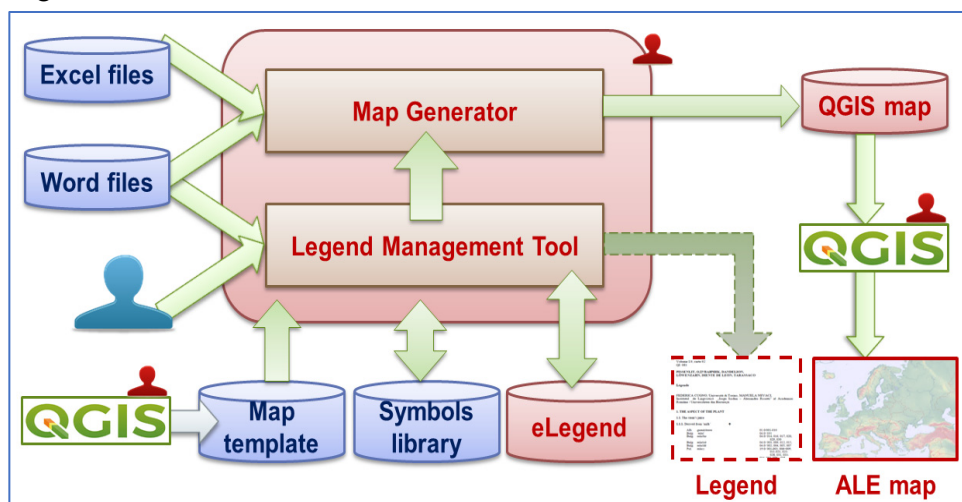


Figure 4. Structure of the GIS based systems for linguistic maps preparation
(source: [Bejinariu *et al.*, 2019])

First, the Map template was created using the QGIS¹⁴ application and the Legend (linguistic database) is created either by editing it from scratch or by importing its content from old legends prepared as Word documents. Then, the map generator automatically creates the linguistic map as a QGIS project by inserting in the map template the symbols, legends, map title, etc. Finally, the QGIS application is used to print the maps or to export them as images or PDF files.

This system was developed by ICS in collaboration with ILIR for the preparation of 9th fascicle of *Atlas Linguarum Europae*¹⁵ [Bejinariu *et al.*, 2018].

¹⁴ <https://www.qgis.org/en/site/>

¹⁵ <https://lingv.ro/atlas-linguarum-europae/>

Also, a new version is being adapted for the preparation of the next volume in *ALRR. Sinteza*. The system is able to create the indexes required for the publication but only if the linguistic database contains all the necessary information.

4 *Framework for Dialectal Data Management*

The framework proposed in this paper is used for the preparation of the *Linguistic Atlas of Romanian Dialects from the North and South of the Danube* (ALDRo). The development of previously implemented systems is imposed by particularities of *ALDRo*. It combines all four dialects of the Romanian Language, some of the investigations took place a long time ago, and another part is to be carried out, on the territory of several Balkan countries. *ALDRo* emphasises the dialectal similarities of North and South-Danubian Romanian dialects (Daco-Romanian, Aromanian, Megleno-Romanian, Istro-Romanian) spoken in Romania, R. Moldova, Ukraine, Serbia, Hungary, Albania, Bulgaria, R. of North Macedonia, Greece and Croatia. The research is based on the questionnaire in *ALE* and the *ALE* dialectal archive from ILIR, in order to highlight the concordances with the Romance languages, with the languages of the Balkan linguistic union and other European languages.

Therefore, in order to meet the mentioned requirements, the *ALDRoMaps* application is being implemented. Even if it is necessary to modify the data structures, *ALDRoMaps* remains compatible with its previous versions due to the implementation of some procedures for the conversion of the data structures used. The structure of the *ALDRoDB* database was inspired by the data management systems used in the previously published atlases, considering that the main information is the same. *ALDRoDB* is an extension which allows the sharing of information included in previous implementations of the linguistic atlases whether they have been computerized or not.

In previous versions the linguistic database was stored locally as a binary file. Although this variant does not allow the editing of linguistic information concurrently by several users, the solution was imposed by the hardware limitations of that time. In order to unify the information edited by several computers, import-export procedures at word level were made available. To eliminate this shortcoming, the new *ALDRoDB* is implemented using the MySQL relational database management system.

The information included in the database can be classified into five categories: general information about the atlases, main linguistic information (related to the words, points, speakers information and phonetic transcriptions), multimedia information (audio, video, images and text), relations definitions (which allow to link the linguistic information to the multimedia files), dynamic values definitions

(for some fields whose list of possible values can change). The structure of the database has been designed to make it possible to specify many answers corresponding to a word and a survey point (Fig. 5).

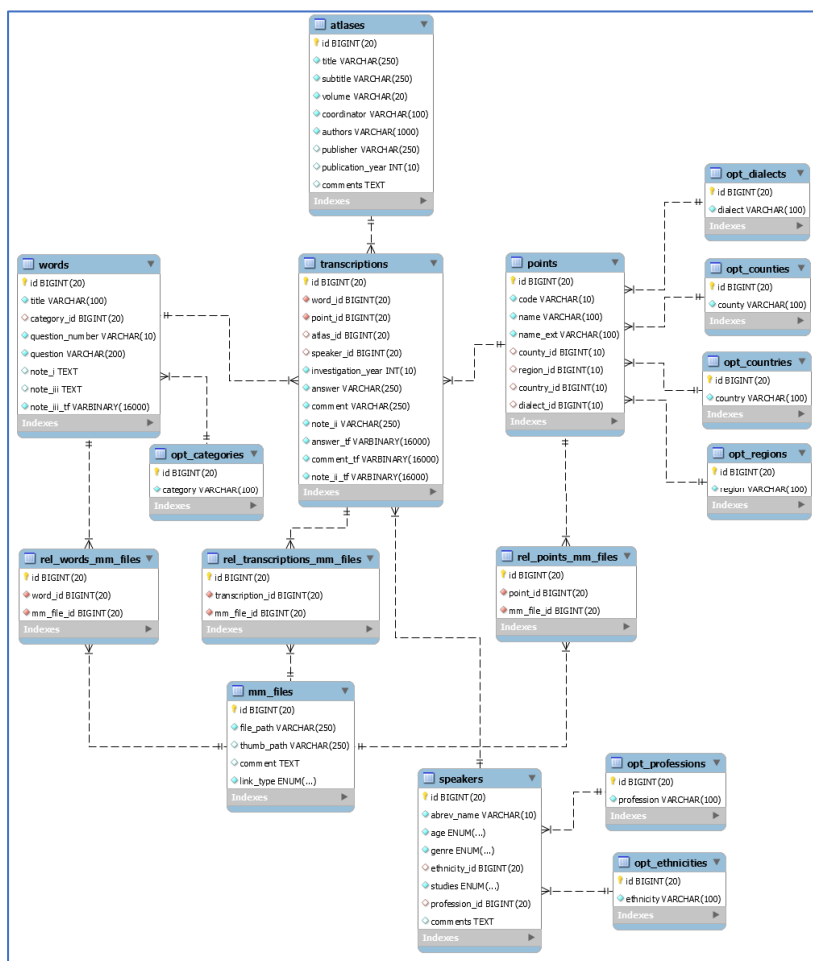


Figure 5. ALDRoDB tables and relationships

4.1 Atlases Information

The 'atlases' table describes the volumes in which the dialectal information was published and it contains fields in which the titles, authors' names, publisher,

publication year, etc. are stored. It makes possible to store multiple atlases content in a single database.

4.2 Main Linguistic Information

This category contains three tables: words, points, speakers and transcriptions. Unlike previous implementations, in ALDRoDB for the fields which contain phonetic transcriptions (*note_iii_tf* in *words* table, *answer_tf*, *note_ii_tf* and *comments_tf* in the *transcriptions* table) were added fields that contain the transliteration of the corresponding phonetic transcriptions. They will be used for operations to search and index the content of the language database. The *speakers* table stores general and non-personal information about the person who gave the answers during the investigation. This allows to the study sociolinguistic variation, by using the diagenational and diasexual variations among the parameters pursued in the surveys [Olariu and Olariu, 2016].

Apart from speaker information, the transcriptions table is also enriched with fields which allow specifying the atlas in which the answer was originally published and the investigation year. This because between the investigation and publishing time can be a difference of decades – which is not a too long time interval for the language evolution study. In the points table were added fields to better identify the investigation points as: county, region or country and also the spoken dialect.

4.3 Multimedia Information

In the previous versions of the linguistic databases there was room for associating an image for words and an audio recording for answers, respectively. In ALDRoDB there is the possibility to associate one or more multimedia files (audio, video, image and text) to each word, each point and each answer in the database. The multimedia files are statically stored on the storage devices and their paths together with their descriptions are stored in the *mm_files* table. The associations between these files and the corresponding words, points and answers are established through the relationships' tables: *rel_words_mm_files*, *rel_points_mm_files* and *rel_transcriptions_mm_files*. Other seven tables (*opt_categories*, *opt_dialects*, *opt_counties*, *opt_regions*, *opt_countries*, *opt_professions* and *opt_ethnicities*) are used to store the dynamic lists of possible values for some fields in the main tables: *category* in table *words*; *dialect*, *county*, *region* and *country* in table *points*; *profession* and *ethnicity* in table *speakers*. Other static lists of possible values of other fields are implemented using the *enum* data type available in SQL.

4.4 *Phonetic Transcriptions*

As it was described in previous works, the phonetic transcriptions are synthesized in real-time. For database consulting and maps visualization from the *ALDRoMaps* application, the transcripts are generated by its internal component. If these operations take place online on the Internet, synthesizing the image of the transcripts is relatively difficult. For this reason, among the multimedia information associated with the words and the phonetic transcriptions include their pre-generated images at a resolution necessary for convenient visualization.

4.5 *Geographical Database*

The geographical database is prepared using the QGIS package. It describes both the map template and the geographic features. It is organized in layers using the shape file format using the Stereo-70 coordinate system which is specific to Romania.

Among the layers that describe the map we mention: map limits, map background, map title, map frame, word title, word notes, word image, phonetic transcriptions and notes related to phonetic transcriptions and custom layers. The map background layer describes the position of a possible background image containing the relief of the studied area. The geographic features are described using the following layers: external borders, maritime borders, major rivers, lakes, capitals, major cities and investigation points. Due to the uneven density of the survey points, the phonetic transcriptions are placed on the linguistic maps in pre-defined positions to avoid their overlapping.

4.6 *ALDRoMaps Application*

The *ALDRoMaps* application is implemented in C++ as Windows application. It contains multiple interfaces for editing the content in each of the main tables of the linguistic database. Another interface is used for the selection of parameters for maps generation and their editing (Fig. 6).

Three main improvements were implemented in *ALDRoMaps* application comparing to its previous versions. The first one is related to the possibility to show a previously generated image as map background. Usually, the image contains the relief of the studied area but it can include also other geographical features as rivers, lakes or cities. The user can change the way the map is displayed (with or without background) at any time. Obviously, this change also involves the need to show or hide certain layers in the map, as those geographical features are included or not in the background image. The second improvement is that the user can add custom layers of predefined typed in the map. The following types of objects can be added: straight or curved lines and polygons, texts and phonetic transcriptions. The custom polygons can be drawn with transparency which allows

showing isoglosses on the phonetic maps. An example of *ALDRo* map is shown in the Fig. 7.

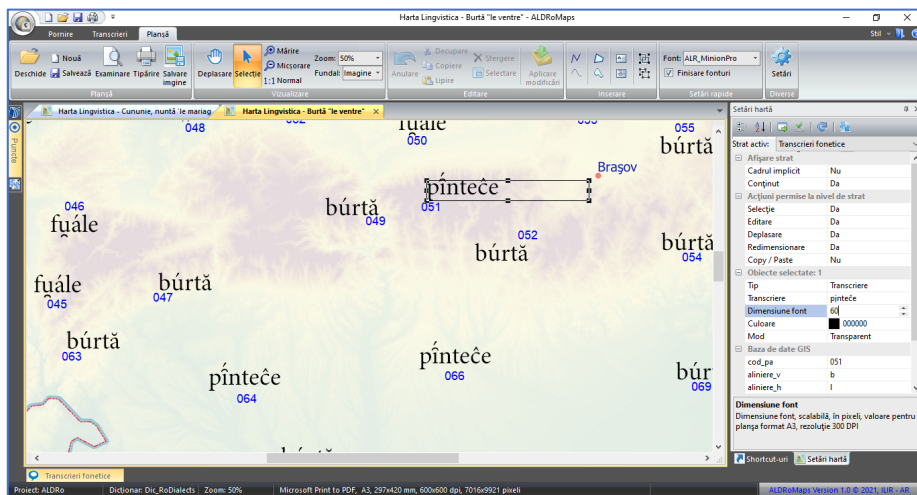


Figure 6. Interface for linguistic maps editing (source: screenshot with the *ALDRoMaps* application implemented by the authors)

The most important improvement implemented in *ALDRoMaps* is that the linguistic content of the maps is loaded from *ALDRoDB* each time it is drawn unlike previous versions in which changes in the linguistic database were not automatically transferred to previously generated maps. Another newly implemented functionality is the possibility to export indexes and lists of words, survey points and phonetic transcriptions in Word files.

5 Conclusions and Future Work

In the last 20 years, important contributions have been made regarding the creation of computerized research tools for Romanian dialectologists. In this paper, a new framework for the Romanian linguistic atlases is proposed. Even if the system is still being implemented, it is already used for the elaboration of the Linguistic Atlas of Romanian Dialects from the North and South of the Danube.

New functionalities of the *ALDRoMaps* application will be implemented in the future. First of all, search functions, automatic generation of word and form indexes as well as other dialect analysis tools will be implemented. It is also necessary to complete the *ALDRoDB* database with the structures needed to define groups of survey points according to the phonetic features used, in order to generate interpretive maps. Last but not least, it is considered the development of

PRESERVATION OF ROMANIAN LINGUISTIC HERITAGE

Web instruments for atlases consulting and the implementation of tools for the integration of socio-linguistic texts in the database.

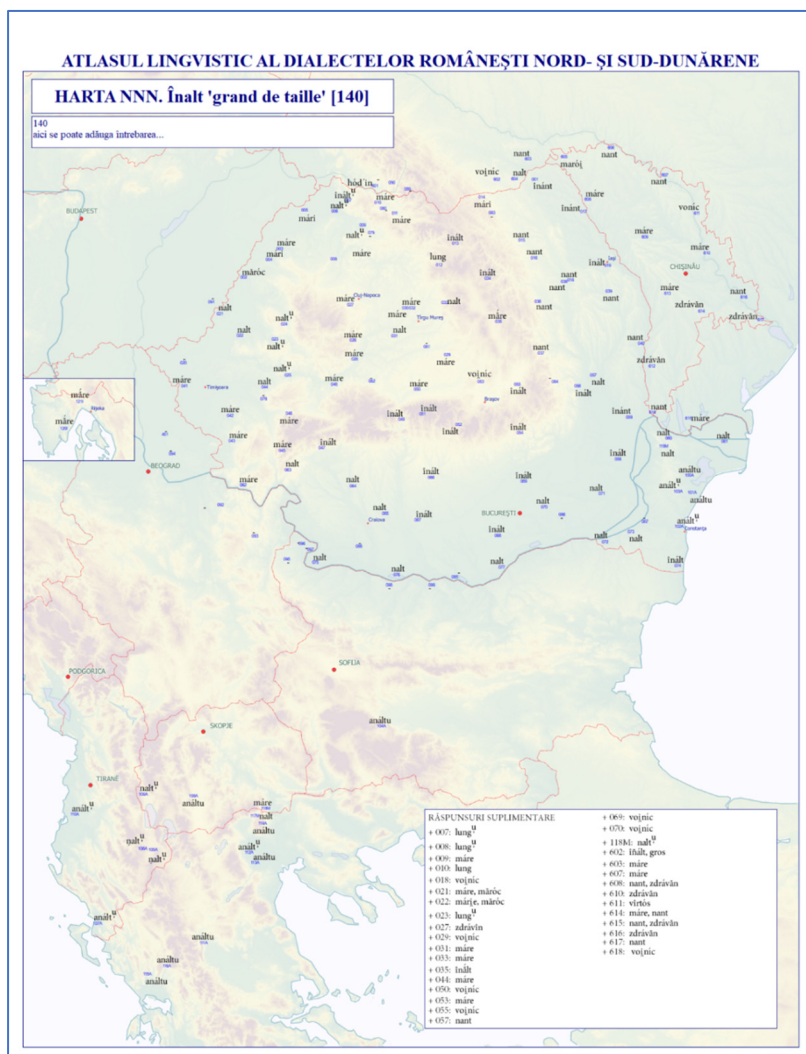


Figure 7. Map in the ALDRo Linguistic atlas (source: image created by authors using ALDRoMaps application)

Comparing to the previous implementations the proposed framework contains some improvements which allow storing in a single database the dialectal information from all the Romanian linguistic atlases.

Acknowledgements

This work presented in this paper was partially supported by the research grant from the Romanian National Authority for Scientific Research and Innovation CNCS = UEFISCDI, project no PN-III-P4-ID-PCE-2020-0451.

References

- Apopei, V., Bejinariu, S.-I., Roman, M. (2002), Graphic Symbols Generator for the Phonetic Transcription in the Electronic Linguistic Atlas. In *Proc. of ECIT'2002, European Conferences on Intelligent Systems and Technologies*, Iași, România, July 2002.
- Bejinariu, S.-I., Apopei, V., Nevaci, M., Saramandu, N. (2019). Information Management Framework for the European Linguistic Atlas. In *Atlas Linguarum Europae (ALE) Symposium, The 52nd Annual Meeting*, Vilnius, Lithuania, 8-12 september 2019.
- Bejinariu, S.-I., Apopei, V., Nevaci, M., Saramandu, N. (2018). New GIS based Approaches for the Linguistic Atlases. In V. Păiș, D. Gifu, D. Trandabăț, D. Cristea, D. Tufiș (eds.) *Proceedings of the 13th International Conference "Linguistic Resources and Tools for Processing Romanian Language"*, Iași, 22 November 2018, Iasi, Romania, pp. 147-158.
- Bejinariu, S.-I., Olariu, F.-T. (2017). Romanian Linguistic Atlases in digital format – A new approach. In *Philologica Jassyensia*, XIII, 1(25): 13-23.
- Bejinariu, S.-I., Apopei, V., Iftene, F. (2021). Romanian geolinguistics in the digital age (in romanian: Geolingvistica românească în era digitala). In: Manuela Nevaci, Irina Floarea, Ioan-Mircea Farcaș (eds.), *Ex Oriente lux. In honorem Nicolae Saramandu*, Alessandria, Edizioni dell'Orso, «La colonna infinita» 14: 163-176.
- Embleton, S., Uritescu, D., Wheeler, E.S. (2013). Defining dialect regions with interpretations: Advancing the multidimensional scaling approach. In *Literary and Linguistic Computing*, 28(1): 13–22, doi: 10.1093/lc/fqs048.
- Hessle, C., Kirk, J. (2020). Digitising Collections of Historical Linguistic Data: The Example of The Linguistic Atlas of Scotland. In *Journal of Data Mining and Digital Humanities*, Episciences.org, *Special Issue on Visualisations in Historical Linguistics*, 1-17, fihal-02166186v2.
- Olariu, F.-T., Olariu, V. (2010). O sută de ani de cartografie lingvistică românească-un bilanț deschis. In *Philologica Jassyensia*, An VI, Nr. 1 (11), pp. 89–118.
- Olariu, F.-T., Olariu, V. (2014). The Romanian linguistic cartography in the digitizing era: the electronic atlases. In *Dialectologia et Geolinguistica*, 22(1): 75-90, doi: 10.1515/dialect-2014-0005.
- Olariu, F.-T., Olariu, V., Luca, R. (2016). Atlasul lingvistic audiovizual al Bucovinei (ALAB) – repere metodologice. In D. Loșonți, V. Vlasin & N. Mocanu (ed.). *Lucrările celui de-al XVI-lea Simpozion internațional de dialectologie*. Cluj-Napoca: Argonaut & Scriptor, pp. 327-329.
- Saramandu, N. (1990). Pentru un atlas al atlaselor lingvistice regionale. In *Limba română*, XXXIX (1), pp. 57-67, 1990.
- Saramandu, N., Nevaci, M. (2013). *Sinteze de dialectologie română*. Editura Universitară, București, pp38-65, doi: 10.5682/9786065917422.

- Saramandu, N., Nevaci, M. (2018). Atlase lingvistice. In Marius Sala, Nicolae Saramandu (coord.), *Lingvistica românească*, Editura Academiei Române. pp: 451-458.
- Scherrer, Y., Samardžić, T., Glaser, E. (2019). Digitising Swiss German: how to process and study a polycentric spoken language. In *Language Resources and Evaluation*, 53, 735–769, doi: 10.1007/s10579-019-09457-5.
- Scherrer, Y., Kellerhals, S. (2014). Digitizing the linguistic atlas of German-speaking Switzerland. In *Methods in Dialectology*, XV, Groningen, <https://archive-ouverte.unige.ch/unige:39955>.
- Sousa, X. (2017). From field notebooks to automatic mapping: the ‘Atlas Lingüístico Galego’ database. In *Dialectologia et Geolinguistica*, 25(1): 1-22, doi: 10.1515/dialect-2017-0001.
- (ALAB II) *Atlasul lingvistic audiovizual al Bucovinei – ALAB II*, <http://www.philippide.ro/alab/rezumati.html>, last accessed on 8.04.2021.
- (ALAR) Nicolae Saramandu, Manuela Nevaci (editor). 2020. Atlasul lingvistic al dialectului aromân, vol. II, Editura Academiei Române, București. <https://lingv.ro/wp-content/uploads/2021/11/Atlasul-lingvistic-al-dialectului-aroman-ALDA.pdf>, last accessed on 10.12.2021.
- (ALD) Linguistic Atlas of Dolomitic Ladinian and neighbouring dialects (“Atlant lingvistic dl ladin dles Dolomites y di dialec vejins”), <https://www.ald.gwi.uni-muenchen.de/en/?db=ald2>, last accessed on 19.11.2021.
- (ALDC) The Linguistic Atlas of the Catalan Domain (L’Atles Lingüístic del Domini Català), Institut d’Estudis Catalans. Barcelona. <https://aldc.espais.iec.cat/>, last accessed on 19.11.2021.
- (ALE) Nicolae Saramandu, Manuela Nevaci and Ionuț Geană (coordinators). 2015. Atlas Linguarum Europae – ALE. Volume I. Neuvième fascicule. Cartes Linguistiques Européennes. București, Editura Universității din București. <https://lingv.ro/atlas-linguarum-europae/>, last accessed on 19.11.2021.
- (ALI) Atlante Linguistico Italiano, <https://www.atlantelinguistico.it/atlante/introduzione.html>, last accessed on 19.11.2021.
- (ALiR) Gaston Tuaillon, Michel Contini. 1996. Atlas Linguistique Roman. Istituto poligrafico e Zecca dello Stato, Libreria dello Stato.
- (ALPI) García Mouton, Pilar (coord.), Inés Fernández-Ordóñez, David Heap, Maria Pilar Perea, João Saramago, Xulio Sousa. 2016. Atlas Lingüístico de la Península Ibérica. CSIC. <http://alpi.csic.es/en>, last accessed on 19.11.2021.
- (ALR) Digital Library of the Institute of Linguistics and Literary History "Sextil Pușcariu", Cluj-Napoca, Atlasul Lingvistic Român, <http://inst-puscariu.ro/alr.html>, last accessed on 10.12.2021.
- (ALRR.Sinteza) Nicolae Saramandu (coord), Mihaela Morcov, Manuela Nevaci, Daniela Răuțu, Irina Floarea, Carmen-Ioana Radu, Ionuț Geană and Mara Iuliana Manta. 2018. Atlasul lingvistic român pe regiuni. Sintează. vol. III, Editura Academiei Române.
- (OLA) General Slavic Linguistic Atlas – OLA. <http://ola.zrc-sazu.si/index.htm>, last accessed on 19.11.2021.
- (WLAD) Linguistischer Atlas des dacorumänischen Sprachgebietes / herausgegeben auf kosten der rumänischen Akademie in Bukarest von Prof. Dr. Gustav Weigand - Leipzig: Johann Ambrosius Barth, 1909, available with Public Domain License at https://www.europeana.eu/en/item/9200456/handle_123456789_54079, last accessed at 07.04.2021.

Old Romanian Lexicons and Their Representation in DLR

MIHAI ALEX MORUZ^{1,2} and MĂDĂLINA UNGUREANU¹

¹*Department of Social Sciences and Humanities,
Institute of Interdisciplinary Research,
"Alexandru Ioan Cuza" University of Iași*

²*Faculty of Computer Science,
"Alexandru Ioan Cuza" University of Iași*

mmoruz@info.uaic.ro, andreea.ungureanu@uaic.ro

Abstract

The expansion of Catholicism and Protestantism in Eastern Europe in the first half of the 17th century resulted in a counter-reform of Orthodoxy, a context that determined the issue of several Slavonic linguistic instruments, among which Pamvo Berynda's Slavonic-Ruthenian Lexicon [Kiev, 1627], the most significant lexicographic work of the time in the Slavonic area. This movement had an equally strong impact upon the Romanian area, mainly in Wallachia, and thus six Slavonic-Romanian dictionaries were written in the second half of the 17th century, having Lex.Ber. as a source, and related in an as yet unknown manner. These lexicons are available as manuscripts and are being digitized within the eRomLex project, which aims at the elaboration of a comparative digital edition.

While analysing these lexicons, we have become interested in the manner in which the lexical inventory from these sources has been integrated into the Thesaurus of the Romanian Language (if at all) from the standpoint of attestation chronology, phonetic or morphologic variants and senses; therefore, we have studied the occurrences of words, glosses and parts of glosses that are to be found in examples and glosses of the DLR.

This paper will describe the methods which we have employed for the search and matching of these digitized Slavonic-Romanian dictionaries to DLR, together with numerical data regarding matches.

1 Introduction

The expansion of Catholicism and Protestantism in Eastern Europe in the first half of the 17th century resulted in a counter-reform of Orthodoxy and reactions from the Orthodox clergy; the best known is that of the Metropolitan of Kiev, Petru Movilă. His activity to increase the prestige of Orthodoxy over Catholicism and Protestant currents was carried out including by writing or translating numerous official or dogmatic works, a context that determined the issue of several Slavonic linguistic instruments, among which Pamvo Berynda's Slavonic-Ruthenian

Lexicon [Kiev, 1627 = Lex.Ber.], the most significant lexicographic work of the time in the Slavonic area. The activity of Petru Movilă also had an impact in the Romanian area, mainly in Wallachia, due to the cultural and economic relations between the two countries. This explains the intense activity of elaboration of linguistic instruments (grammars and lexicons) which took as models works elaborated in Ukraine (the most famous being Meletie Smotrițki's grammar, issued in 1619, and Lex.Ber.). Both the grammar and the lexicon were well-known to the Romanian scholars of the time, due to the close relation between the Romanian and Ukrainian cultures, a relationship that reaches its climax during the reign of Petru Movilă, who supported the foundation of a printing house in Târgoviște by sending material aid and qualified workers; along with the existing economic, political and social factors, this relation of Romanian-Ukrainian cooperation contributed to the extension of the period in which the Slavonic bias dominated the Romanian culture [Cazacu 1984, 211; Panaitescu 1965, 192]. Pamvo Berynda's lexicon was one of the books used to teach the Slavonic language in the college founded by Petru Movilă in Kiev [Ševčenko 1984, 22], which explains its use in Wallachia and its success among the Romanian scholars, along with Meletie Smotrițki's Grammar.

The first lexicographic instruments appeared in the Romanian language, as it happened in other cultural areas, as a result of the diglossia characterizing the culture on the Romanian territories in the old period. Thus, the coexistence of a language of culture, the Slavonic language, and a vulgar language that claimed to acquire the statute of the former, along with the fundamental differences between the two languages led to an increasing need for bilingual lexicographic instruments that could facilitate the cultural transfer between the two languages coming into contact. As the grammar and the dictionary mentioned above were the most famous and used in the Romanian area, it is only natural that they were taken as models for compiling Romanian linguistic tools.

Six Slavonic-Romanian dictionaries were elaborated on the Romanian territory having Lex.Ber. as a source (directly or indirectly, an aspect that is still to be clarified). They have not been fully studied so far, but, besides the fact that they come from the same source, five of them, at least, seem to come from the same archetype, which was lost or has not been found. Their group is impressive by the dimensions of the lexicons, by the connections between them, by the fact that they all seem to come from the same geographical area (Wallachia). Assessing the place of this dictionary group in the broader framework of the Romanian lexicographical activity during the 17th century, it is obvious that they constitute its most consistent part. This led us to the idea of a comparative edition; in fact, this idea was issued more than 100 years ago by the editor of the oldest of these lexicons (see [Crețu, 1900]), who thought that a comparative edition might be a useful tool for getting an answer to the problem of the links between them.

The eRomLex project¹, which is being developed at the "Alexandru Ioan Cuza" University in Iași, financed by a grant of the Ministry of Research, Innovation and Digitization, aims at providing a comparative edition of the first Romanian bilingual lexicons, starting from the fact that they have the same source, in order to highlight the characteristics of this lexicographical network (the affiliations between lexicons, the way they relate to the source, the innovations towards it, their probable uses) and to facilitate the access to their content. Also, we thought of a digital edition because it would allow exhaustive data extraction and comparison and link with other digitized dictionaries and resources for old Romanian or Church Slavonic, including dictionaries. After a first stage, of filing (transcription into Latin alphabet of the definitions from the Romanian lexicons and transcription of the Slavonic headwords both from Lex.Ber. and from the Romanian ones, respecting the graphic features of each lexicon), follows the stage of automatic processing and alignment of entries from the Romanian lexicons having as pivot the corresponding entry from Lex.Ber. Alignment is based on the degree of similarity of the headwords and on the location of the entries in the lexicons. In the case of headwords that are not in Lex.Ber., the entries in Romanian lexicons are aligned with each other, according to the same principles.

The comparative edition will be a valuable lexicographic resource for the dictionaries of old Romanian (including the Thesaurus of the Romanian language - DLR), for specialists in the history of the literary Romanian language, historical dialectology and translation studies, providing new data with regard to this specific period of the old Romanian culture. Moreover, the outcomes will provide good opportunities for the study of old equivalence strategies specific to early Romanian lexicography, as well as for the study of the characteristics of these types of dictionaries and their circulation at the time.

This paper will investigate means by which these old Romanian lexicons, which are now being edited and stored in a structured format, could be used as tools for the creation of new lexical resources for the Romanian language, or to aid in the extension and improved coverage of existing lexical resources, such as the DLR.

2 The Corpus: the Slavonic-Romanian Dictionaries from the 17th Century

Our corpus consists of the six Slavonic-Romanian lexicons created on the Romanian territory having Lex.Ber. as a source, preserved until the present day. All these lexicons are manuscripts; four of them can be found at the Romanian Academy Library (B.A.R.), one belongs to the Russian National Library of Sankt Petersburg while another is held at the Russian State Archive of Old Documents

¹ See <http://www.scriptadacoromanica.ro/bin/view/eRomLex/> and also Gînsac *et al.* (2021).

in Moscow. All these works date from the second half of the 17th century (except for Lex.Mard., 1649).

1. Lex.Mard. = The Lexicon of Mardarie Cozianul from Rom. ms. no. 450 from the National Library of Romania in Bucharest, 1649; 200 leaves; edited by Gr. Crețu [1900];

2. Lex.St. = The Lexicon from. Rom. ms. no. 312 from the National Library of Romania in Bucharest (41r-216v) attributed to Staicu Grămăticul [Strungaru, 1966] or to Daniil Panoneanul [Ursu, 2003]; followed by an Old Slavonic-Romanian grammar, translation of Smotrițki's grammar mentioned above; dated: 1667-1669 [Strungaru, 1966] or cca 1660-1670 [Mihăilă 1972, 313];

3. Lex.1348 = The Lexicon from Rom. ms. no. 1348 from the National Library of Romania in Bucharest (1-84v), signed by someone named Mihai Logofătul; dated 1683 [Ștrempel 1978, 300] or 1678 [Mihăilă 1972, 313];

4. Lex.3473 = The Lexicon from Rom. ms. no. 3473 from the National Library of Romania in Bucharest (pages 1-369), written by someone named Mihai (1672-1673); followed by a Slavonic-Romanian grammar; the files containing the entries beginning with A are missing; see Gînsac and Ungureanu [2021];

5. Lex.Mosc. = The Lexicon from Moscow (held at the Russian State Archive of Old Documents, Fond 188, Оп. 1. Ч. 2., p. 491); 282 leaves; the title page is missing; see Ciobanu [1914], Gînsac and Ungureanu [2019].

6. Lex.Pet. = The Lexicon from Petersburg (Lexicon ce să zice cuvinte pe scurt alese din limba slovenească pre limba rumânească), 100 leaves; held at the National Library of Russia in Sankt Petersburg (notice n° Q.XVI.5 – Славяно-молдавский словарь, 100 листов, составлен в 1695 году); dated around 1693 [Crețu 1900, 50] or ante 1693 [Mihăilă 1972, 315]; see Bogdan [1981].

7. Lex.Ber. = Рамво Berynda, Леґиконъ славенороскїй и именъ тлъкованїе, Kiev, 1927 [2nd edition: Kutein, 1653]. Online: <http://litopys.org.ua/berlex/be.htm>.

Since these 6 dictionaries have not been edited thus far, except for Lex.Mard., their coverage in the Romanian Thesaurus Dictionary (Dicționarul Limbii Române – DLR) is low. Although the Lex.Mard. edition is part of the bibliography of some Romanian dictionaries, such as DLR and Tiktin, references to another lexicon, Lex.St., were made using the manuscript directly, and, therefore, contain errors (for example, for the first known written use of the noun “sârguială”, en. “dilligence”, Tiktin references Lex.St., 200 – however, the word cannot be found either on page 200v or on the page 200r, or on some other pages we have checked manually); the other lexicons are not used at all. Given the fact that these six dictionaries represent the largest part of Romanian lexicography in the 17th century, and, with the exception of Anonymus Caransebesiensis, are the very first dictionaries for the Romanian language, we can assume that their editing and storage in an electronic format will provide important resources for the Romanian

Thesaurus with regard to first attestations, unrecorded terms, phonetic variations or meanings that have never been attested.

3 *Electronic Comparative Edition*

The electronic version of the comparative edition is conceived in such a way as to allow for complex interrogation, extraction and exhaustive comparison of the data, in a manner which a typical, printed edition, could never allow. There is, however, the possible drawback that a digital version provides access to the content of the dictionaries but is not conducive to the viewing of the dictionaries in a manner close to the original manuscript format; in other words, the digital version allows for an efficient use of the linguistic content, but also obscures the original presentation of the text, the display of the marginal notes, etc. Our priority, however, is to be able to examine the linguistic content in order to provide answers to the questions and hypotheses put forward until now, or, more accurately, to gain a better understanding of this lexicographical network and the scholars who contributed to it. Upon completion of the digital version, we will provide the means by which the lexical information will be exported and used for the creation of standalone, printed versions of these dictionaries.

In order to obtain the electronic comparative edition of these 6 dictionaries, together with their alignment, we have used as a pivot the progenitor of the Romanian versions, Lex.Ber. Since these lexicons are bilingual (Slavonic to Romanian), the original word from Lex.Ber. is used as the headword (sometimes in a slightly modified form of writing, given the vagaries of handwriting on a very expensive medium, e.g., *Бавилѡнъ* in Lex.Ber., *бавилѡ* in Lex.St., or even *бавѡлѡ* in Lex.Mard.). Because of this, during the manual transcription phase, we have manually indicated the exact entry in Lex.Ber. to which the entry in the Romanian lexicon is aligned to, and, on this basis, all the equivalent entries are grouped together. Also, to improve alignment, especially in the case of those words which do not appear in Lex.Ber., we have applied a transformation to the Slavonic form in order to unify the written representation (replacement of overwritten letters with their inline equivalent, unification of diacritics, reduction to the same letter in the case of vowels, etc.).

Source Volume	Head Word	Page	Column	Number	Gloss
Lex.Ber.	Бѣръцѣ		16	14	
Lex.Pet.	Бѣръцѣ	9v	I	16	sunare. clopoțelu
Lex.Staicu	Бѣръцѣ	50v	I	14	drângăire, sunare, clopoțe/
Lex.Mosc.	Бѣръцѣ	20v		13	sunare clopoțe/
Lex.Mard.	Бѣръцѣ	32		3	coarde sau ciocâne/
Lex.3473	Бѣръцѣ	11v		14	sunare, clopoțe/
Lex.1348	Бѣръцѣ	6r	II	20	sunare clopoțe/

Figure 1 Aligned entries for the word Бѣръцѣ

The aligned entries are useful from a twofold perspective: firstly, we have access to various manners of translating the same Slavonic word, and such variation can prove interesting from a lexicographic and historical standpoint; secondly, there are some cases where the same Lex.Ber. entry is referenced in two (or even more) distinct entries in the same dictionary. Since the entries are not given in strict alphabetical order, it has been quite difficult to notice such occurrences until now. Below, we have given a figure showing the aligned entries for the word Бѣръцѣ "tympanum; organum" (Miklosich, s.v.).

Source Volume	Head Word	Page	Column	Number	Gloss
Lex.Ber.	ѡмѣвънѣ		341	5	
Lex.Pet.	ѡмѣвънѣ	2r	II	10	loc mai înalt puținel
Lex.Staicu	ѡмѣвънѣ ѡмѣвънѣ	45v, 42v	I	6	scaun de poveaste loc înalt
Lex.Mosc.	ѡмѣвънѣ	4v		11	loc mai înalt puținel
Lex.Mard.	ѡмѣвънѣ	8		1	scaun de poveaste
Lex.3473					
Lex.1348	ѡмѣвънѣ	2r	I	10	loc mai înalt puținel

Figure 2 Aligned entries for the word ѡмѣвънѣ. Note the entries in Lex.St.

The columns in Fig. 1 give the source volume (as a sigle), the form of the head word as written in the respective lexicon, information regarding position in the

volume (page, column if applicable, and a number indicating the position of the entry in that column) and the associated translation (gloss). The last column contains notes regarding transcription or other philological observations, where applicable.

In Fig. 2 we have given an example of multiple Romanian equivalents given for the same Lex.Ber. entry, in different entries in the same volume. In this particular case, in Lex.St., the entry *Ѧмвѡнѣ*² is given both on page 42v and on page 45v, with different glosses (it is worth noting that both glosses can be found separately in entries from other dictionaries): one is "loc înalt" (= "elevated place"); the other is "scaun de poveaste", meaning "chair or space intended for discussions" (see DLR, scaun (3), but also (5), (6); poveste (6), but rather the meaning referring to 'discuție, sfat' – 'discussion', which does not appear separately in DLR).

As can be seen from the examples given above, the alignment of the six lexicons yields important lexical information with regards to the circulation of Romanian words and word senses in the 17th century. Also, alignment contributes to the clarification of the sources and relations between the lexicons.

4 *Matching the Lexicons with DLR*

While carrying out the work of aligning and analyzing these lexicons, we have become interested in the manner in which the lexical inventory from these sources has been integrated into the DLR (if at all) from the standpoint of attestation chronology, phonetic or morphologic variants and senses; therefore, we have begun to study the occurrences of words, glosses and parts of glosses that are to be found in examples and glosses of the DLR.

In order to carry out these investigations, we had to take into account the peculiarities of the Romanian equivalents given in the lexicons. The types of glosses seem to conform to three broad categories: single word equivalents (*ѡрава* – "alăută" – "harp"), sequences of synonyms (*ѡѡспѣтїѡнїѡ* – "uspătătoriu, cîrciumar" – "waiter, inkeeper") or explanatory glosses (*Ѧмвѡнѣ* – "aceasta iaste pîiatră scumpă și cine o poartă la sine nu-l prinde vinul niciodată" – "this is a gemstone and whoever wears it will never get drunk"). For the purposes of our current study, we have disregarded the third type of definition (explanatory glosses), as we were chiefly interested in matches denoting word equivalence; we intend to carry out analysis regarding gloss equivalence at a later date.

² The translation of the gloss from Lex.Ber.: "Elevated place. In relation to certain meetings: a chair or circle (= space) for talking (= discussions). No. [this is the short form for the Book of Neemias] 8:4. (2Ez. 9:42)". The Ruthenian definitions were translated by Ion-Mihai Felea and Cecilia Maticiu, within the eRomLex project.

The analysis was carried out on part of the lexicons (the Slavonic letters **Ѧ**, **Ѧ** and **Ѧ**³). Firstly, we have automatically extracted those entries which contained at least one aligned gloss of the first two types (single words or sequences of synonyms). For those glosses containing sequences of synonyms (actually, lists of words separated by commas or other punctuation), we have automatically separated the words, and thus reduced the problem to searching for single words. The words extracted were then used as queries to automatically search the eDTLR (the electronic version of the DLR); these searches were carried out on two levels: matches on the title word and searches in examples.

In the case of searches over the headwords, the purpose of the searches was to find entries in the DLR which had the same headword as the search term; since the query terms are usually archaic variants, we did not attempt a perfect match, but went for a similarity match instead (experiments have shown that an 80%-word similarity provides the most relevant results). This search returned a set of potentially equivalent entries, which would then need to be manually examined in order to determine semantic equivalence or relevance. For example, for the query “uspătătoriu”, the results returned were the entries “OSPĂTĂTOR, - OĂRE”, “OSPĂTĂTORIU, - OĂRE” and “USPĂTĂTORIU, -OĂRE”. Upon manual examination of these results, we have determined that the last two are variants of the first one; upon further examination of the entry “OSPĂTĂTOR, - OĂRE”, and considering the glosses provided in all of the six lexicons (in particular the gloss from Lex.St. “ospătător, a toŭ[i] priimitor”), it might be justified to add one more example to sense 1 of the DLR entry (synonym to “Ospitalier”).

The second type of search we have carried out was performed over all the examples in the eDTLR; in this case we have only taken into account perfect matches, as we were interested in determining the extent of the coverage of the exact form taken into analysis with regards to DLR examples. For example, in the case of the word **ВЛѦХВОРАЊІЕ** – “vrăjire”, “sorcery”, we have extracted 8 entries which contain examples using the word, three of which have headwords which were returned in the title word search as well.

In both methods employed, however, the search results that we have obtained automatically cannot be used without human analysis, but rather serve to point the experts towards those particular cases of interest which might be used to further enrich, extend or even improve the DLR. Therefore, a qualitative evaluation of these results is, by its very nature, time consuming.

³ The eRomLex Project uses the font CyrillicaBulgarian10U kindly provided by prof. Anna-Mariya Totomanova, from “St. Kliment Ohridski” University of Sofia, leader of the project „Cyrillomethodiana” (<https://cyrillomethodiana.uni-sofia.bg/>).

5 *Matching the Lexicons with DLR*

In this paper we have described a method of employing some of the oldest known Romanian dictionaries in order to improve existing lexical resources, with a case study on the Romanian Thesaurus Dictionary. We have shown that, by providing automatically generated matches, we could aid lexicographers in extending or improving existing lexical resources by providing examples of use for words from the second half of the 17th century. The eRomLex corpus is useful for the DLR for the clarification of the senses of some terms in the context of the old Romanian language, firstly, by association with the Slavonic headword, which has a Romanian gloss, and, secondly, by means of the Ruthenian definition, translated by the authors of the lexicons (see the above discussion for “scaun de poveaste”, where the sense of the word “poveste” is more likely that of “talk, discussion, conference”, which is not included in DLR).

For further research, we intend to improve the interrogation procedure by taking into account descriptive glosses as well; this could be done either by attempting to match glosses in the DLR directly to glosses in the lexicons, by means of multiword searches, or by searching for phrases in those DLR word senses preceded by diamond makers and written in bold or italic font.

Acknowledgements

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P1-1.1-TE - 2019-0517, within PNCDI III.

References

- Bogdan, I. (1981). Un lexicon slavo-român din secolul XVII. *Convorbiri literare*, 25:193-204.
- Cazacu, M. (1984). Pierre Mohyla (Petru Movilă) et la Roumanie: Essai historique et bibliographique. *Harvard Ukrainian Studies*, 8(1/2): *The Kiev Mohyla Academy: Commemorating the 30th Anniversary of its Founding (1632)*:188-222.
- Сіобану, Ș. (1914). Славяно-румынскій словарь бібліотеки Московскаго Общества Истории и Древностей No 240. *Русскій филологическій вестникъ*, 71(1):75-88.
- Crețu, G. (ed.). (1900). *Mardarie Cozianul. Lexicon slavo-românesc și tâlcuirea numelor din 1649*. Edițiunea Academiei Române, Institutul de Arte Grafice “Carol Göbl” București.
- Academia Română. 2010. *Dicționarul limbii române*. Ediție anastatică, Editura Academiei, București.
- Gînsac, A.-M., Moruz, M.-A., Ungureanu, M. (2021). Slavonic-Romanian Lexicons of the 17th Century and their Comparative Digital Edition (the eRomLex Project). *Diacronia*, 14:1-11.
- Gînsac, A.-M., Ungureanu, M. (2019). Lexiconul de la Moscova în raport cu celelalte lexicoane slavo-române din secolul al XVII-lea. *Philologica Jassyensia*, 30:245-258.

- Mihăilă, G. (1972). Contribuții la studiul lexicografiei slavo-române din secolul al XVII-lea. In *Contribuții la istoria culturii și literaturii române vechi*, Editura Minerva, București, pages 307-324.
- Miklosich, F. (1865). *Lexicon Palaeoslovenico-Graeco-Latinum*. Braumueller, Vindobonae (online: <http://monumentaserbica.branatomic.com/mikl2/main.php>).
- Panaiteescu, P.P. (1965). *Începuturile și biruința scrisului în limba română*. Editura Academiei Române, București.
- Ševčenko, I. (1984). The Many Worlds of Peter Mohyla. *Harvard Ukrainian Studies*, 8(1/2): *The Kiev Mohyla Academy: Commemorating the 30th Anniversary of its Founding (1632)*: 9-44.
- Strungaru, D. (1966). Începuturile lexicografiei române. *Romanoslavica*, 13:141-158.
- Ștrempel, G. (1978-1992). *Catalogul manuscriselor românești*, volume 1 (1978), volume 2 (1983), volume 3 (1987), volume IV (1992). Editura Științifică și Enciclopedică, București.
- Tiktin, H. (2000-2005). *Rumänisch-deutsches Wörterbuch*. 3. Auflage, hrsg. von Paul Miron und Elsa Lüder. Harrassowitz, Cluj-Napoca/Wiesbaden.
- Ursu, N.A. (2003). Contribuții la istoria culturii românești în secolul al XVII-lea. Studii filologice. Editura Cronica, Iași.

CHAPTER 3

DATASETS AND CORPORA

RoITD: Romanian IT Question Answering Dataset

DRAGOȘ CONSTANTIN NICOLAE and DAN TUFIȘ

Research Institute for AI Romanian Academy

{constantin, tufis}@racai.ro

Abstract

Recent improvements in the language modeling field have advanced the latest state-of-the-art performance for a variety of natural language processing tasks. Reading comprehension in particular has seen dramatic progression. Most frameworks are based on the English language with only a few limited frameworks available in other languages. In this paper, we introduce a Romanian Question Answering Dataset focused on IT and electronics topics known as Romanian IT Dataset (RoITD). RoITD is a reading comprehension dataset in the Romanian language consisting of 9575 QA samples. The baselines for this framework have been trained using Multilingual BERT, Multilingual DistilBERT, and Romanian BERT. The Exact Match is reported to be 35.48%, 34.15%, 35.06% and F1 score of 50.94%, 45.72%, 53.62% for Multilingual BERT, Multilingual DistilBERT, and Romanian BERT respectively.

1 Introduction

Advancements in language modeling have resulted in many successful results for a variety of natural language processing (NLP) tasks. Machine reading comprehension (MRC) is a NLP task requiring computers to comprehend a text and then respond to relevant questions. MRC is vital in the development of a powerful search engine and intelligent agent framework, but their training datasets are expensive, difficult to gather, and almost exclusively based on the English language. These kinds of datasets (e.g., SQuAD1.1 [Rajpurkar *et al.*, 2016], SQuAD2.0 [Rajpurkar *et al.*, 2018], and CoQA [Reddy *et al.*, 2019]) have been used to dramatically improve English Question Answering models. Similar types of datasets based on other languages include the Chinese dataset of the span-extraction MRC [Cui *et al.*, 2019], Chinese user-query-log-based dataset of DuReader [He *et al.*, 2018], Korean MRC dataset [Lim *et al.*, 2019], and French Question Answer dataset (FQuAD) [d’Hoffschmidt *et al.*, 2020]. However, the development of alternative language Question Answering models continues to lag behind English-based models due to the scarcity of annotated datasets in other languages, especially Romanian.

In order to complete alternative language QA tasks, we introduce a Romanian IT Dataset (RoITD) resembling SQuAD 1.1. RoITD consists of 9575 Romanian QA pairs formulated by crowd workers. QA pairs are based on 5043 articles from

Romanian Wikipedia articles describing IT and household products. Of the total number of questions, 5103 are possible (i.e., the correct answer can be found within the paragraph) and 4472 are not possible (i.e., the given answer is a "plausible answer" and not correct). A RoITD sample is provided in Fig. 1.

To evaluate the dataset performance, the QA model was fine-tuned using available transformer-based pretrained models: Multilingual BERT [Devlin *et al.*, 2019b], DistilBERT [Sanh *et al.*, 2020], and A Romanian BERT Model (RoBERT) [Masala *et al.*, 2020]. Although deep learning models, especially transformer-based models, have surpassed the human accuracy level on MRC tasks, it is still unknown how these models will perform in other languages.

The primary contributions of this paper are as follows:

- A benchmark dataset for MRC in native Romanian consisting of 9575 QA pairs based on 5043 articles is introduced. The dataset is freely available for research purposes.

```

"version": "v1.0",
"data": [
  {
    "title": "IT1",
    "paragraphs": [
      {
        "qas": [
          {
            "question": "Ce fel de telefon este iPhone SE",
            "id": "RM53H9975",
            "answers": [
              {
                "text": "cel mai puternic telefon de 4 inch din toate timpurile",
                "answer_start": 25
              }
            ],
            "is_impossible": 0
          }
        ]
      }
    ],
    "context": "Iti prezentam iPhone SE, cel mai puternic telefon de 4 inch din toate timpurile. Pentru a-l crea, Apple a pornit de la un design de mare succes, apoi a reinventat tot restul. A9 este acelasi cip avansat folosit in iPhone 6s. Camera de 12 megapixeli face poze incredibile si filmeaza 4K, iar functia Live Photos iti aduce imaginile la viata. Rezultatul este un iPhone care pare mic, dar traieste pe picior mare. In design mult iubit, iPhone SE preia un design incredibil de popular si il aduce la un nou nivel de rafinament."
  }
]

```

Figure 1 A QA pair example in RoITD.

- The data is evaluated using several linguistic techniques including length-based analysis (i.e. question length, answer length, and passage length) and type-based analysis (i.e. question type, answer type, and reasoning type).
- The first experimental MRC evaluation of RoITD was conducted using Romanian BERT and multilingual BERT, as well as their variations.

Related datasets in various languages along with SQuAD are presented in Section 2.

2 Related Datasets

Aside from English, native reading comprehension datasets are scarce. A few have been built using the same method as SQuAD1.1 in Chinese, Korean, Russian, and French. The SberQuAD dataset [Efimov *et al.*, 2020] is a Russian native reading comprehension dataset containing over 50,000 items. The KorQuAD dataset [Lim *et al.*, 2019] is a 70,000-plus sample size Korean native reading comprehension dataset. Additionally, the CMRC 2018 dataset [Cui *et al.*, 2019] is a Chinese-native reading comprehension dataset that includes over 20,000 QA pairs.

Recently, a French MRC dataset has been created with almost 65,000 QA samples [d’Hoffschmidt *et al.*, 2020].

These datasets are used to test and develop deep neural network natural language processing models such as Match-LSTM [Wang and Jiang, 2017], BiDAF [Seo *et al.*, 2017], R-Net [Wang *et al.*, 2017], DrQA [Chen *et al.*, 2017], FusionNet [Huang *et al.*, 2018], FastQA [Weissenborn *et al.*, 2017], and QANet [Yu *et al.*, 2018]. BERT [Devlin *et al.*, 2019a] and XLM-R [Conneau *et al.*, 2020a] are powerful models trained on many languages that have recently achieved state-of-the-art results on MRC datasets.

Since language-specific datasets are expensive and difficult to come by, cross-lingual models can offer a potential solution. Cross-lingual models can be used to transfer existing datasets to a desired language without the need for additional training data in the target language [Lewis *et al.*, 2020]. These unsupervised multilingual models have been demonstrated to generalize successfully in zero-shot cross-lingual scenarios [Artetxe *et al.*, 2020]. As a result, crosslingual question answering has increasingly grown more popular, leading to the development of two cross-lingual benchmarks, XQuAD [Artetxe *et al.*, 2020] and MLQA [Lewis *et al.*, 2020]. The MLQA dataset contains over 12,000 QA samples in English, as well as 5,000 samples in Arabic, German, and Spanish. Another option is to translate a dataset into the target language, then fine-tune a language model using the translated data. Carrino *et al.*, used this technique to develop Translate Align Retrieve (TAR), a unique translation approach, to convert the English SQuAD1.1 dataset to Spanish [Carrino *et al.*, 2020]. The translated SQuAD1.1 dataset has been subsequently used to fine-tune a multilingual model. However, this kind of translation is not always reliable and is highly dependent on the pretrained multilingual model. In response to these issues, we introduce a MRC task RoITD, a native Romanian language-based dataset consisting of 5492 QA pairs from 2224 Wikipedia articles derived by crowd-sourcing.

3 Dataset Analysis

3.1 Dataset Collection

The Romanian-language QA dataset (RoITD) focuses on IT and electrical themes. It is comprised of 5492 QA pairs created by crowd workers. The QA sets are based on 2224 Romanian Wikipedia articles about IT and household items. Of the total number of questions, 3480 are possible (i.e., the correct answer can be found in the paragraph) and 2012 are not possible (i.e., the answer given is a "plausible answer" and is not correct). Unlike other QA data sets, this dataset contains the question, answer, and passage containing the full context of the QA pair, including the corresponding response order in order to meet the planned architecture of BERT-special. The machine completing the QA assignment must be able to read natural language text and comprehend (i.e., use) its content to extract relevant

information from the main paragraph in the same manner that people do when answering questions.

The image shows a web form titled "Add Custom Question". It features a text input field for "Question text" containing the text "Ce fel de telefon este iPhone SE|". Below this is a dropdown menu for "Question category" with the letter "U" selected. A blue "Submit" button is at the bottom.

Figure 2 Annotation process.

Crowd workers were required to create up to four questions for each provided paragraph where at least one question was unanswerable. Impossible questions were solely based on the provided paragraph in such a manner that crowd workers could identify a feasible response within the text. Furthermore, crowd workers were instructed that inquiries should not always be structured using the exact phrasing found in the context, although all questions were required to be grammatically correct. Crowd workers were then requested to combine the query with the relevant category in the following way: If the question is unanswerable and the answer is likely, use the letter "U", and if the question is answerable and the answer is correct, use the letter "A". Fig 2 shows the user interface for this annotation process while Fig 3 shows the crowd worker confirmation after the correct registration of a question. The user interface used by the crowd workers was hosted using Amazon Web Service t2.micro machine with an Ubuntu 20.04.1 operating system.

3.2 Dataset Structure

The data structure follows the format of SQuAD, which contains several attributes such as "question", "id", "text", "answer_start", "is_impossible", and "context". The paragraph provided to crowd sourcing workers is stored in the field "context". This incorporates manually-selected paragraphs from Wikipedia. The field "id" is

comprised of a randomly assigned unique identification number for the answer-question pair. Only the numbers "0" and "1" are allowed in the "is impossible" field. The category "A" is assigned the value "0", indicating that the answer is correct. The value "1" corresponds to the category "U", indicating a plausible answer. The question posed by the source crowd source worker is represented by the field "question". The field "answer_start" keeps track of the character index marking the beginning of an answer.

The screenshot displays the RoITD interface. On the left, under the 'Questions' tab, a question is listed: 'Ce fel de telefon este iPhone SE' with a status of 'U' and a unique ID of '1'. Below this is a button labeled 'ADD CUSTOM QUESTION'. On the right, the 'Annotation Document' tab is active, showing a Romanian text passage about the iPhone SE. A search bar is located at the top right of the document area.

Figure 3 After the question is formulated.

3.3 Dataset Categorization

Fifty questions were randomly selected from the dataset, and the instances were manually categorized using the categories indicated in Table 1. This process provided a stronger understanding of the required reasoning needed to answer questions as categorized by SQuAD [Rajpurkar *et al.*, 2016]. The results demonstrate that all cases in the passage exhibit a lexical or syntactic difference between the question and the answer. Additionally, 100 negative examples were manually analysed to gain a deeper understand of their potential issues. The categorization outlined by Rajpurkar *et al.*, [2018] was used in this study, as shown in Table 6. A wide range of occurrences beyond the expected are distinguished, such as negation, antonym, and entity changes. The results of this analysis show that 98% of the sampled negative cases were unanswerable.

3.4 Dataset Statistics

Table 1 Categorization of 50 selected samples based on reasoning.

Reasoning	Description	Example	Percentage
Lexical variation	Major correlation between the question and the answer sentence are synonyms	Q:Ce se poate detecta cu ajutorul camerelor duale AI? Sentence: Cu camerele duale AI din spate de 13 MP + 2 MP, se pot recunoaste cu usurinta obiecte ...	54%
Lexical variation(world knowledge)	Major correlation between the question and answer sentence requires world knowledge to resolve.	Q:Cu ce este asemanator Evelatus? Sentence: ...un telefon similar cu Evelatus Easy ...	12%
Multiple sentence reasoning	There is anaphora, or higherlevel fusion of multiple sentences is required	Q:Ce ajuta sa faci selfie-uri pe timp de noapte? Sentence: ... Nokia 230 are o camera frontala de 2 MP cu blitz cu LED si o carcasa din aluminiu subtire si eleganta.... Blitz-ul cu LED de pe partea frontala a telefonului te ajuta sa faci selfie-uri pe timp de noapte	25%
Ambiguous	we do not agree with the crowd-worker's answer, or the question does not have a unique answer	Q:De ce V4 are ergonomie marita? Sentence: ... datorita formei speciale si a finetii date de colturile rotunjite, V4 Viper iti ofera o experienta placuta la atingere ...	9%

Table 3 provides a general summary regarding the RoITD dataset size. The table shows that the dataset is split into train and test sets. Other relevant information includes the total number of articles and questions, average question and answer length, and vocabulary sizes.

Table 2 Types of unanswerable questions in RoITD

Reasoning	Description	Example	Percentage
Negation	Negation word inserted or removed.	S: ...Filmul se remarca prin realismul stiintific precum si prin efectele speciale folosite in premiera.) Q: Prin ce nu se remarca filmul?	65%
Antonym	Antonym used.	S: ... Camera principala de 64MP surprinde clar momentele preferate, iar lentilele cu unghi larg, lentila macro si senzorul de adancime va vor asigura ca veti obtine cele mai bune rezultat Q: Ce asigura lentilele cu unghi ingust?	12%
Entity Swap	Entity, number, or date replaced with other entity, number or date	S: Consola contine 2 butoane de tip analog stick (spre deosebire de predecesorul sau care detine unul), un ecran OLED multi-tactil cu o diagonala de 5inch (12,7cm), de asemenea are Bluetooth, Wi-Fi si optional 3G. Intern are un procesor 4 core ARM Cortex-A9 core si o placa grafica SGX543MP4+. Q: Ce functii are butonul suplimentar al consolei?	8%
Other Neutral	Other cases where the paragraph does not imply any answer and	S: ... Trei camere care actioneaza ca una.Grandangular 13 mm. Unghi de camp de 120. Scena de patru ori mai mare Q:Ce greutate are Apple iPhone 11 Pro?	15%
Answerable	Question is answerable (i.e., dataset noise).	-	0%

3.5 Length-Based Analysis

It is important to analyse the size- and length-based statistics of any given dataset. This analysis has been applied to the RoITD dataset including diversity analysis of article, question, and answer sizes. The length based on word count for all three categories is shown in Table 4, 5, and 6. Table 4 shows that the majority of questions fall into the 6–10-word length range, and very few question samples exceeding 15 words are present within the dataset. Intuitively speaking, this short question length is to be expected. Similarly, Table 5 shows a similar distribution

within the dataset with regard to answer length where the majority of answers fall into either the 1-5 or 6-10-word count range. Lastly, Table 6 shows that the majority of passages consist of the word counts ranging from 51 to 80 (i.e., 46.23% of passages in train and 62.33% of passages in test). Passages exceeding 80 words were very rare in both the train and test data.

Table 3 Types of unanswerable questions in RoITD

	Train	Test	All
Number of articles	4170	813	5043
Number of questions	7175	2400	9575
Average passage length	52.72	61.97	55.04
Average question length	8.08	8.12	8.11
Average answer length	13.44	7.85	9.25
Vocabulary size	38265	18396	48821

3.6 Type-Based Analysis

Table 4 Question length statistics in RoITD

Length	Train	Test
1-5	16.86	18.33
6-10	64.04	62.54
11-15	17.42	16.25
16-20	1.57	2.25
> 20	0.09	0.12

The dataset was also analysed with respect to question type. To complete the analysis, inquiries were divided into the following eight categories: Who, What, When, Where, Why, How, How many, and Others. The Romanian version of these question are: Care, Ce, Cand, Unde, De, Cum, Cati, and Alti. The crowd worker manually selected a question type to ensure diversity within the dataset. The linguistic structure of Romanian is quite similar to English. However, the portion of What-type questions in RoITD is lower than that in SQuAD (53.60%). The question type distribution is shown in Fig. 4. Questions starting with “What” (i.e., “Ce”) are dominant in the dataset. This is due to the fact that RoITD consists of IT based questions pertaining to equipment, which typically start with “What”. “How many” (i.e., “Câți”) is the second most dominant question type followed by “How” (i.e., “Cum”) and “Who” (i.e., “Care”).

4 Evaluation

To evaluate RoITD, tests were ran using state-of-the-art MRC models. Human performance was also evaluated using the RoITD MRC problem to assess the dataset difficulty. Two evaluation metrics, exact match (EM) and F1-score, were implemented to evaluate the MRC model performance using RoITD, similar to evaluations used for English [Rajpurkar *et al.*, 2016] and Chinese [Cui *et al.*, 2019] datasets.

4.1 Implementation Using State-of-the-Art Model

Various BERT-based models were run using RoITD to establish baselines for further improvement. Multilingual pretrained forms of various transformer-based models were used to complete this task. The specific models include the following:

- BERT-multilingual: Devlin *et al.* [2019b] proposed the BERT method. This model is a powerful tool for pre-training language representations and has shown cutting-edge outcomes in a variety of reading comprehension tests. For this evaluation, a pretrained large-scale multilingual language

Table 5 Answer length statistics in RoITD

Length	Train	Test
1-5	48.62	19.45
6-10	23.13	23.95
11-15	14.16	21.62
16-20	7.67	16.29
> 20	6.39	18.66

Table 6 Passage length statistics in RoITD

Length	Train	Test
1-30	10.71	3.58
31-50	37.12	25.04
51-80	46.23	62.33
81-100	5.70	6.75
> 100	0.22	2.29

model, mBERT, was used.

- Romanian BERT: Romanian BERT is a BERT model trained entirely on a 15GB Romanian corpus collected by cleaning OSCAR, OPUS, and Wikipedia sub-corpora [Dumitrescu *et al.*, 2020]. This model easily outperforms the only available multilingual BERT, mBERT, on various NLP tasks.
- DistilBERT: DistilBERT is a general purpose version of BERT that is 40% smaller and 60% faster, although it still retains almost 97% of language comprehension accuracy [Sanh *et al.*, 2020]. DistilBERT demonstrates the capability for on-device calculations using both proof-of-concept experiments and comparative on-device analysis.

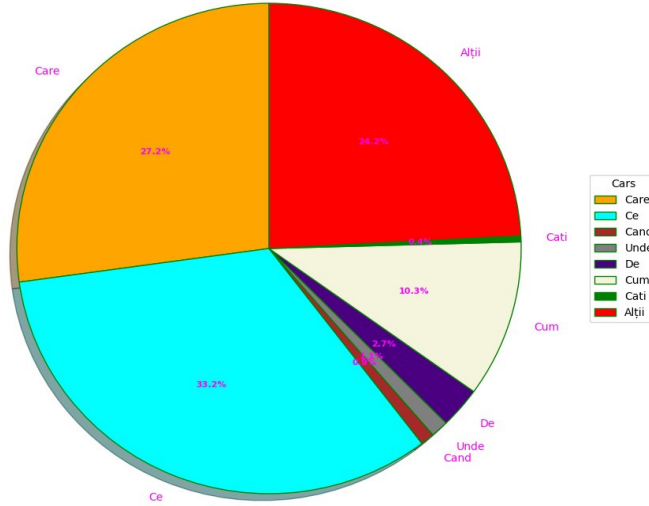


Figure 4 The distribution of the question types

- XLM-R: XLM-R is a powerful methodology for pretraining scalable multilingual language models, resulting in considerable increases in a variety of cross-lingual transfer tasks [Conneau *et al.*, 2020b]. On a range of crosslingual benchmarks such as XNLI, MLQA, and NER, this model surpasses multilingual BERT (mBERT). XLM-RBase and XLM-RLarge were tested on the RoITD dataset in this study.

Table 7 Baseline performance for RoITD dataset

Models	Exact Match (EM)	F1-Score
bert-base-multilingual-uncased (mBERT)	35.48	50.94
bert-base-romanian-uncased-v1 (romBERT)	35.06	53.62
distilbert-base-multilingual-cased (distilBERT)	34.15	45.72

4.2 Experimental Results

To train all MRC models on RoITD, a single NVIDIA Tesla P100 GPU via Google Collaboratory was used. The configuration for BERT-based pretrained models was constant for all experiments. To provide a proper comparison, all of the selected models were initialized using the same parameters. The specific parameters were as follows: batch size per gpu = 12, maximum sequence length

= 384, document stride = 128, and thread number = 4. The learning rate was $3e^{-5}$ and the training duration was 10 epochs. The performance was evaluated based on the standard framework (i.e., Exact Match (EM) and F1-score). The result of each BERT-based model is shown in Table 7. BERT-base-Romanian-uncased easily outperforms mBERT and DistilBERT with an EM of 35.06% and F1-score of 53.62. This result is to be expected since romBERT is specifically trained in the Romanian language with a specific vocabulary set that covers semantics compared with mBERT. Even though the F1-score is higher than mBERT, the EM is almost the same for both models. The lower performance of DistilBERT can be attributed to the fact that it is a smaller model compared with romBERT and mBERT.

5 Conclusion

In this paper, we introduce a new MRC task for evaluating Romanian datasets related to IT and electronic inquiries. The proposed RoITD dataset consists of 9575 question-answer pairs generated by crowd source workers. To the best of our knowledge, this is the first native Romanian language based MRC task specifically related to the IT and electronics domain. The designed dataset is diverse due to its varied question types. In addition, the dataset performance was evaluated using state-of-the-art models such as BERT and XLM. The baseline reported in this paper shows superior performance for BERT models specifically trained in the Romanian vocabulary. However, there is still significant room for improvement as both the EM and F1 scores are below 35% and 55%, respectively. In the future, we intend to increase the size and quality of RoITD to improve deep learning and transformer model performance. Furthermore, we would like to open the Romanian MRC challenging task to other researchers in the field.

References

- Artetxe, M., Ruder, S., Yogatama, D. (2020). On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online, July. Association for Computational Linguistics.
- Carrino, C. P., Costa-jussà, M., Fonollosa, J. (2020). Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France, May.
- Chen, D., Fisch, A., Weston, J., Bordes, A. (2017). Reading wikipedia to answer open-domain questions. In *ACL*.
- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020a). Unsupervised crosslingual representation learning at scale. In *ACL*.

- Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V. (2020b). Unsupervised crosslingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.
- Cui, Y., Liu, T., Che, W., Xiao, L., Chen, Z., Ma, W., Wang, S., Hu, G. (2019). A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5883–5889, Hong Kong, China, November. Association for Computational Linguistics.
- Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.
- Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019b). 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, Minneapolis, Minnesota, June. Association for Computational Linguistics.
- d’Hoffschmidt, M., Belblidia, W., Heinrich, Q., Brendlé, T., Vidal, M. (2020). FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online, November. Association for Computational Linguistics.
- Dumitrescu, S., Avram, A-M., Pyysalo, S. (2020). The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328, Online, November. Association for Computational Linguistics.
- Efimov, P., Boytsov, L., Braslavski, P. (2020). Sberquad - russian reading comprehension dataset: Description and analysis. In *CLEF*.
- He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T., Wang, H. (2018). DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia, July. Association for Computational Linguistics.
- Huang, H., Zhu, C., Shen, Y., Chen, W. (2018). Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *ArXiv*, abs/1711.07341.
- Lewis, P., Oguz, B., Rinott, R., Riedel, S., Schwenk, H. (2020). MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July. Association for Computational Linguistics.
- Lim, S., Kim, M., Lee, J. (2019). Korquad1.0: Korean qa dataset for machine reading comprehension.
- Masala, M., Ruseti, S., Dascalu, M. (2020). RoBERT – a Romanian BERT model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6626–6637, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.
- Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical*

- Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.
- Rajpurkar, P., Jia, R., Liang, P. (2018). Know what you don’t know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July. Association for Computational Linguistics.
- Reddy, S., Chen, D., Manning, C. (2019). CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, March.
- Sanh, V., Debut, L., Chaumond, J., Wolf, T. (2020). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.
- Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H. (2017). Bidirectional attention flow for machine comprehension. *ArXiv*, abs/1611.01603.
- Wang, S., Jiang, J. (2017). Machine comprehension using match-lstm and answer pointer. *ArXiv*, abs/1608.07905.
- Wang, W., Yang, N., Wei, F., Chang, B., Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *ACL*.
- Weissenborn, D., Wiese, G., Seiffe, L. (2017). Fastqa: A simple and efficient neural architecture for question answering. *ArXiv*, abs/1703.04816.
- Yu, A., Dohan, D., Luong, M., Zhao, R., Chen, K., Norouzi, M., Le, Q. (2018). Qanet: Combining local convolution with global self-attention for reading comprehension. *ArXiv*, abs/1804.09541.

A Puzzle-Based Dataset for Natural Language Inference

ROXANA SZOMIU and ADRIAN GROZA

*Department of Computer Science,
Technical University of Cluj-Napoca, Cluj-Napoca, Romania*
{roxana.szomiu, adrian.groza}@cs.utcluj.ro

Abstract

We provide here a dataset for tasks related to natural language understanding and natural language inference. The dataset contains logical puzzles in natural language from three domains: comparing puzzles, knights and knaves, and zebra puzzles. Each puzzle is associated with the entire set of atomic questions that can be generated based on the relations and individuals occurring in the text. For each question we provide the correct answer: entailment, contradiction or ambiguity. The answer's correctness is verified against theorem provers. Good puzzles have two properties: (i) each piece of information is necessary and (ii) no unnecessary information is provided. These properties make puzzles interesting candidates for machine comprehension tasks.

1 Introduction

Recognising textual entailment (RTE) is a classical problem in natural language processing (NLP), that aims to identify the relationship between sentence pairs, specifically if they entail or contradict each other, or neither of those. RTE has a wide range of applications including question answering, spam detection, sentiment analysis.

Datasets play an important role in assessing the RTE systems by evaluating their ability to predict the correct label-relationship for two sentences. Building a new dataset can be a difficult task because texts can be extremely rich source of information, but if they are unstructured, extracting information from them can be challenging and time-consuming.

The identification of correct relation - entailment, contradiction, unknown - can be performed either through manual annotation or by automatic labelling. Our proposed dataset consists of various puzzle, and based on each puzzle's text we generate all possible questions and then using logical theorem provers, we provide the answers for each question. The proposed dataset contributes in evaluating Question Answering applications. Most of the recent advances in question answering (QA) proposed deep learning solutions and the progress was made in open-domain question answering as proved by Minaee *et al.* [2021].

Our contribution is the creation of dataset for textual entailment based on different types of logic puzzles. The dataset contains 355 unambiguous puzzles with 4,136 questions, and 401 ambiguous puzzles with 16,745 questions. Starting from the puzzle’s text, the generated FOL theory is taken by the Mace4/Prover9 to automatically assess the entailment between puzzle’s text and the newly generated question. Solving puzzles in first order logic can be a challenging task too. To the best of our knowledge, there is no puzzle-based dataset designed to be used for Recognizing Textual Entailment, that contains the entire set of all possible questions, that can be answered from the initial text. In the absence of a similar dataset, our approach based on Question-Answering technique, aims to generate question/answer pairs from the text of puzzles.

2 Datasets for Inference Tasks

This section briefly describes the available datasets that are widely used for textual entailment or question answering task.

The Sentence Involving Compositional Knowledge (SICK) is a large dataset of human intuitions on English sentences, collected through crowdsourcing by Marelli *et al.* [2014]. SICK includes about 10,000 sentence pairs, each annotated for the degree of semantic relatedness and the type of entailment relation: entailment, contradiction, and neutral. The entailment annotation led to 5,595 neutral pairs, 1,424 contradiction pairs, and 2,821 entailment pairs. The SICK dataset was constructed starting from two existing sets: the 8K ImageFlickr¹ and the SemEval 2012 STS MSR-Video Description².

The Excitement-Open-Platform (EOP) described by Magnini *et al.* [2014], relies on various datasets such as RTE-3 English data set, Excitement dataset³ provided by Kotlerman *et al.* [2015], SICK⁴, described in Marelli *et al.* [2014], OMQ⁵, which is a RTE-style dataset, semi-automatically created from manually categorized German customer requests. EOP consists of 800 English text-hypothesis (T-H) pairs for training and 800 T-H pairs for testing. Each pair is annotated with one of the three classes: entailment, non-entailment or unknown. Also, each pair is labelled with the appropriate text inference task: information extraction (IE), information retrieval (IR), question answering (QA), or summarization (SUM).

The Recognizing Textual Entailment⁶ (RTE) datasets proposed by Dagan *et al.* [2005] come from a series of textual entailment challenges and are constructed

¹ <http://nlp.cs.illinois.edu/HockenmaierGroup/data.html>

² <http://www.cs.york.ac.uk/semeval-2012/task6/index.php?id=data>

³ <http://hlt-nlp.fbk.eu/technologies/textual-entailment-graph-dataset>

⁴ <https://zenodo.org/record/2787612#.YF86KRKxU5k>

⁵ <https://www.dfki.de/~neumann/resources/omqdata.html>

⁶ <https://tac.nist.gov/2011/RTE/index.html>

based on news and Wikipedia text. The PASCAL RTE datasets have been annotated for contradiction. They are marked for a 3-way decision in terms of entailment: "yes" (entails), "no" (contradicts), and "unknown" (doesn't entail but is not a contradiction). The datasets are not balanced since contradictions represent about 10% of the data.

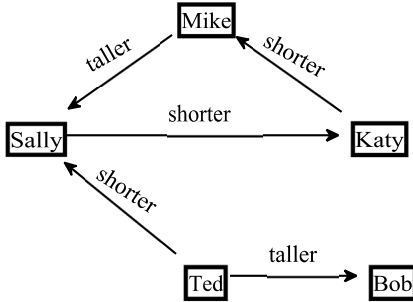
The Guardian Headlines Entailment Training Dataset consists of around 32,000 pairs of sentences (16,233 for which entailment does hold and 16,249 for which it doesn't) automatically extracted from The Guardian using the provided API. Since constructing manually entailment datasets is a time-consuming task, so, they are fairly small, which means machine learning approaches perform sub-optimally. As Hickl *et al.* [2005] showed, automatically constructed datasets can improve the performance of systems using machine learning by up to ten percent.

The Textual Entailment Graph was created within the Excitement project by Kotlerman *et al.* [2015] as a gold standard data to evaluate the task of automatic Textual Entailment Graph (TEG) generation. The main difference between this task and the traditional RTE task is that the text pairs are not independent. The nodes in the graph are inter-connected via entailment edges, which should not represent contradicting decisions. For instance, if the edges (u,v) and (v,w) are in the graph, then the edge (u,w) is implied by transitivity. The English dataset contains a text collection generated on the basis of 389 emails sent by customers of a railway company. Text fragments contain the customers feedback and are clustered into 29, for a total of 756 nodes and 7,830 edges.

Williams *et al.* [2018] have introduced the Multi-Genre Natural Language Inference (MNLI)⁷ containing 433k sentence pairs annotated with textual entailment labels. MNLI is based on the Stanford Natural Language Inference (SNLI) described by Bowman *et al.* [2015]. MNLI covers a range of genres of spoken and written text, and supports a distinctive cross-genre generalization evaluation. SNLI is a collection of 570k human-written English sentence pairs manually labelled for balanced classification with the label's entailment, contradiction, and neutral.

The SciTail dataset is the first entailment dataset created for the science question answering task by Khot *et al.* [2018]. Each premise-hypothesis pair is annotated as entails or neutral. About 43.3% of the questions did not have a single supporting sentence, indicating that these questions either need multiple sentences for question answering or better retrieval results. From the remaining 56.7% (i.e. 1,834 questions), the dataset contains 27,026 examples (10,101 with 'entails' label, 16,925 with neutral label) divided into train/dev/test splits with 23,596/1,304/2,126 examples.

⁷ <https://www.nyu.edu/projects/bowman/multinli/>



(a) Information extracted from Puzzle 1

Question	Theorem in FOL	Answer
Is Mike the tallest ?	<i>tallest(mike)</i>	Entailment
Is Bob the shortest ?	<i>shortest(bob)</i>	Entailment
Is Katy taller than Bob ?	<i>taller(katy,bob)</i>	Entailment
Is Bob shorter than Ted ?	<i>shorter(bob,ted)</i>	Entailment
Is Sally the tallest ?	<i>tallest(sally)</i>	Contradiction
Is Katy the shortest ?	<i>shortest(katy)</i>	Contradiction
Is Bob taller than Katy ?	<i>taller(bob,katy)</i>	Contradiction
Is Mike shorter than Ted?	<i>shorter(mike,ted)</i>	Contradiction

(b) Sample of atomic question with 4 predicates and 5 individuals

Figure 1 Question answering for unambiguous comparison puzzles

Wikipedia has been a source for building QA datasets. Thorne *et al.* [2018] have proposed the Fact Extraction and VERification (FEVER), a dataset that contains 185,445 claims manually verified against the sections from Wikipedia pages and classified as supported, refuted, or not-enough-info. Yang *et al.* [2015] have developed the WikiQA dataset. It consists of a set of question-answer pairs, collected and annotated for open-domain question answer research. The corpus includes 3,047 questions and 29,258 sentences, where 1,473 sentences were labelled as answer sentences to their corresponding questions. The dataset also includes questions for which there is no correct answer, allowing researchers to evaluate answer triggering models. The Stanford Question Answering Dataset (SQuAD) proposed by Rajpurkar *et al.* [2016] is a collection of question-answer pairs extracted from Wikipedia articles. In SQuAD, the correct answers of questions can be any sequence of tokens in the given text. The questions and

answers are produced by crowdsourcing, so it is more diverse than some other question-answering datasets. The first version of SQuAD contains 107,785 question-answer pairs on 536 articles, while SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 un-answerable questions written adversarially by crowdworkers in forms that are similar to the answerable ones.

3 Question Answering on Puzzles

Our puzzle-based dataset for textual entailment (PuzzTE) contains three types of puzzles: (i) comparison type puzzles, (ii) knight and knave puzzles, and (iii) zebra puzzles. The method used to build the dataset relies on natural language processing to automatically identify the characters from each puzzle and relationships among them. Based on manually defined grammar rules we obtain a formal representation of each puzzle in First Order Logic (FOL). The puzzles were solved using Mace4 model finder and Prover9 theorem prover developed by McCune [2005]. We automatically generated all possible atomic questions against each puzzle and compute the correct answer based on automated deduction. Hence, each QA pair is automatically annotated with three labels: (1) *entailment* when a proof is found, (2) *contradiction* when the opposite sentence is proved, and (3) *unknown* in case of not enough information in the puzzle, signalled by the Mace4 tool with more than one interpretation model. Given specific grammar rules for each domain, the same automatic is applied for each type of puzzle exemplified here: comparison, knights and knaves, and zebra puzzles.

3.1 Comparison Type Puzzles

A comparison puzzle describes a scenario that involves ordering relationships. Take the example:

Puzzle 1 *Mike is taller than Sally who is shorter than Katy. Ted is taller than Bob but shorter than Sally. Katy is shorter than Mike. Who is the tallest? Is Katy taller than Bob? Is Mike shorter than Ted?*

By using the grammar rules, the Puzzle 1, represented in Figure 1a, is automatically translated into FOL with three clauses: (1) *taller(Mike,Sally) $\wedge$ shorter(Sally,Katy)*, (2) *taller(Ted,Bob) $\wedge$ shorter(Ted,Sally)*, respectively (3) *shorter(Katy,Mike)*. The manually defined grammar rules in NLTK (Listing 1) are generic for all puzzles from this comparison puzzle domain. The sequences of words are grouped into chunks. An example of a chunk is the noun phrase chunk, which can be noun phrases such as "the girl" or "the boy", or proper names such as "Mike" or "Sally". Unlike the words in lexical rules, which can be considered terminals, the chunks are nonterminals. The linguistic categories have different properties (e.g. plural for nouns). One can specify the number for a noun with the feature NUM: NP[NUM=sg, SEM=?subj]. For verb phrases, the tense or person

can be specified: VP[NUM=sg, TNS=pres, PERS=third]. These features help to assess the correctness of sentences. For instance, the context free grammar in Listing 1 can parse sentences like "Mike is taller than Sally" or "Mike is the tallest", but won't be able to parse the sentence "Katy are the tallest".

```

S [SEM = <?subj (?vp)>] -> NP [NUM=?n, SEM=?subj] VP [NUM=?n, SEM=?vp]

NP [LOC=?l, NUM=?n, SEM=?np] -> PropN [LOC=?l, NUM=?n, SEM=?np]
    | InterPrn [LOC=?l, NUM=?n, SEM=?np]
    | Nom [NUM=?n, SEM=?np]

NP [+COP, NUM=?n, SEM=?nom] -> Det [NUM=?n] Nom [NUM=?n, SEM=?nom]

NP [NUM=?n, SEM=<?det (?adj, ?nom)>] ->
    Det [NUM=sg, SEM=?det] Adj [GRD=?abs, SEM=?adj]
Nom [NUM=sg, SEM=?nom]

VP [NUM=?n, SEM=<?v (?prd)>] -> AuxP [+COP, NUM=?n, SEM=?v]
Pred [SEM=?prd]

VP [NUM=?n, SEM=<?cc (?prd1, ?prd2)>] ->
    VP [NUM=?n, SEM=?prd1] CC [BOUND=?vpToVp, SEM=?cc]
    VP [NUM=?n, SEM=?prd2]

VP [NUM=?n, SEM=<?cc (?prd1, ?prd2)>] ->
    VP [NUM=?n, SEM=?prd1] CC [BOUND=?vpToVp, SEM=?cc]
    AdjP [GRD=?rel, SEM=?prd2]

AdjP [GRD=?rel, SEM=<?quant (?adj (?pp))>] ->
    Quantifier [SEM=?quant] Adj [GRD=?rel, SEM=?adj]
PP [+than, SEM=?pp]

AdjP [GRD=?rel, SEM=<?adj (?pp)>] ->
    Adj [GRD=?rel, SEM=?adj] PP [+than, SEM=?pp]

AdjP [GRD=?rel, SEM=<?neg (?adj, ?pp)>] ->
    Neg [SEM=?neg] Adj [GRD=?rel, SEM=?adj] PP [+than, SEM=?pp]

AdjP [GRD=?abs, SEM=<?adj>] -> Det [NUM=sg] Adj [GRD=?abs, SEM=?adj]

PP [+than, SEM=?np] -> P [+than] NP [SEM=?np]

AuxP [COP=?c, NUM=?n, SEM=<?neg (?aux)>] ->
    Aux [COP=?c, NUM=?n, SEM=?aux] Neg [SEM=?neg]

```

Listing 1 Sample of grammar rules for Puzzle 1

The lexical rules in Listing 2 contain the linguistic categories for each comparison puzzle. Each word belongs to such a category and can have one or more features as Wagner [2010] presented. Two such features used by the discourse checker module are SEM, which specifies the semantic of the word and NUM which defines the number of a noun (sg-singular or pl-plural). We used the

GRD feature that handle the gradable adjectives (e.g., relative gradable adjective: *shorter* or *taller*, or absolute gradable adjective: *shortest* or *tallest*).

Based on the obtained FOL theory, Mace4 computes the models satisfying the given relation. For Puzzle 1, Mace4 generates a single model - the first one from Figure 2. By querying this model, we can automatically generate entailment or non-entailments answers. The set of atomic questions is generated based on the number of individuals from puzzle and the relations between. For the Puzzle 1, there are two unary predicates: *shortest(x)* and *tallest(x)* and two binary predicates: *taller(x,y)* and *shorter(x,y)*, and 5 individuals: *Mike, Sally, Katy, Bob and Ted*. For each puzzle we generated a number of $a * n + b * n^2$ questions, where a is the number of unary predicates, b the number of binary predicates, and n the number of individuals in the given puzzle. For this puzzle, we generated $2*5+2*5^2 = 60$ questions. Among them 22 will generate entailments, and 38 will generate contradictions. A sample of these 60 questions appears in Table 1b.

```

PropN [-LOC, NUM=sg, SEM=<\P.P(Mike)>] -> 'Mike'
PropN [-LOC, NUM=sg, SEM=<\P.P(Sally)>] -> 'Sally'
PropN [-LOC, NUM=sg, SEM=<\P.P(Katy)>] -> 'Katy'
PropN [-LOC, NUM=sg, SEM=<\P.P(Ted)>] -> 'Ted'
PropN [-LOC, NUM=sg, SEM=<\P.P(Bob)>] -> 'Bob'
Det [NUM=sg, SEM=<\P Q x.(P(x) & Q(x))>] -> 'the'
InterPron [-LOC, NUM=sg, SEM=<\P.P(x)>] -> 'who' | 'Who'
RelPron [SEM=<\Q x P.((P(x) & Q(x)))>] -> 'who'
Aux [+COP, NUM=sg, SEM=<\P x.P(x)>, tns=pres] -> 'is' | 'Is'
Aux [+COP, NUM=, SEM=<\P x.P(x)>, tns=pres] -> 'are' | 'Are'
CC [BOUND=?vpToVp, SEM=<\P Q x.((P(x) & Q(x)))>] -> 'but'
Adj [GRD=rel, SEM=<\X x.X(\y.(taller(x,y)))>] -> 'taller'
Adj [GRD=rel, SEM=<\X x.X(\y.(shorter(x,y)))>] -> 'shorter'
Adj [GRD=abs, SEM=<\x.(shortest(x))>] -> 'shortest'
Adj [GRD=abs, SEM=<\x.(tallest(x))>] -> 'tallest'
Quantifier [SEM=<\P. P>] -> 'little'
Quantifier [SEM=<\P. P>] -> 'too'
Question [] -> '?'

```

Listing 2 Lexical rules for Puzzle 1

To extend the dataset with more questions, two approaches can be used: (1) to consider other predicates based on some background knowledge, and (2) to generate puzzles with ambiguity by removing a part of information from the initial puzzle.

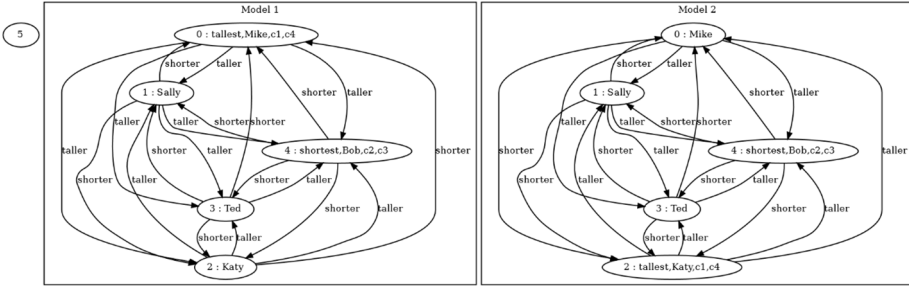


Figure 2 Models computed by Mace4 for Puzzle 1 and Puzzle 2

First, additional predicates can be obtained from some domain knowledge base. For the comparison puzzle domain one binary predicate can be *sameTallAs*(x,y) or *sameShortAs*(x,y). Some additional questions could be: "Are Katy as tall as Mike?" translated to FOL: *sameTallAs*(*Katy*,*Mike*)) or "Are Ted not as short as Bob?" translated to FOL: $\neg$ *sameShortAs*(*Ted*,*Bob*)).

Second, we introduced ambiguity by removing sentences from puzzles. We used the term "ambiguous" for puzzles that are not completely defined: the case when a part of information is missing that prevents us to find a single model. For ambiguous puzzles, computes several interpretations which are models of the input formulas. To generate ambiguous puzzles, we remove sentences (i.e., clues) from the complete puzzles. For instance, by removing *Katy is shorter than Mike* from Puzzle 1, we obtain the ambiguous Puzzle 2.

Puzzle 2 *Mike is taller than Sally who is shorter than Katy. Ted is taller than Bob but shorter than Sally. Who is the tallest? Is Katy taller than Bob? Is Mike shorter than Ted?*

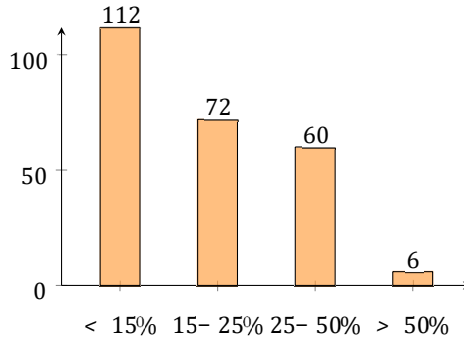
Because of the missing information, one cannot infer whether *Katy is shorter than Mike* or if *Katy is the tallest*. Mace4 signals this ambiguity by computing two models for Puzzle 2 presented in Figure 2.

We quantify the ambiguity of each puzzle by the number of *unknown* answers. For Puzzle 2, the generated FOL theory contains two unary predicates: *shortest*(x) and *tallest*(x), two binary predicates: *taller*(x,y) and *shorter*(x,y), and 5 individuals: *Mike*, *Sally*, *Katy*, *Bob* and *Ted*. Note that by removing some sentences, some predicates or even individuals may disappear from the puzzle. Among all 60 questions, there are now 19 entailment answers, 35 contradictions, and 6 unknown answers. A sample of these 60 questions are listed in Table 3a.

A PUZZLE-BASED DATASET FOR NATURAL LANGUAGE INFERENCE

Question	Theorem in FOL	Answer
Is Katy taller than Sally ?	<i>taller(Katy,Sally)</i>	Entailment
Is Mike taller than Sally ?	<i>taller(Mike,Sally)</i>	Entailment
Is Mike taller than Katy ?	<i>taller(Mike,Katy)</i>	Unknown
Is Katy taller than Mike ?	<i>taller(Katy,Mike)</i>	Unknown
Is Katy the tallest ?	<i>tallest(Katy)</i>	Unknown
Is Mike the tallest ?	<i>tallest(Mike)</i>	Unknown
Is Sally taller than Katy ?	<i>taller(Sally,Katy)</i>	Contradiction
Is Mike shorter than Ted?	<i>shorter(mike,ted)</i>	Contradiction

(a) Sample of atomic questions for ambiguous puzzles



(b) Distribution of ambiguity

Figure 3 Question answering for ambiguous comparison puzzles

By removing one statement or two statements from each puzzle, we can play with the level of ambiguity within the PuzzTE dataset. The level of ambiguity is affected also by the information containing in each sentence. For instance, the statement "Mike is taller than Sally who is shorter than Katy" contains more information than "Katy is shorter than Mike". Hence, the PuzzTE dataset contains puzzles with different levels of ambiguity. The picture 3b shows a histogram with the distribution of ambiguity in puzzles: there are 112 puzzles that have level of ambiguity less than 15 %, there are 72 puzzles with level of ambiguity between 15-25 %, there are 60 puzzles with level of ambiguity between 25-50 %, and only 6 puzzles which have level of ambiguity greater than 50%.

Table 1 Question answering for Knights and Knaves puzzles

Question	Theorem in FOL	Answer
Is Sue a knight ?	$knight(Sue)$	Entailment
Is Bart a knave ?	$knave(Bart)$	Entailment
Is Rex a knight ?	$knight(Rex)$	Entailment
Is Dave a knave ?	$knave(Dave)$	Entailment
Is Sue a knave ?	$knave(Sue)$	Contradiction
Is Bart a knight ?	$knight(Bart)$	Contradiction
Is Rex a knave ?	$knave(Rex)$	Contradiction
Is Dave a knight ?	$knight(Dave)$	Contradiction

The PuzzTE dataset contains 300 comparison type puzzles: 50 puzzles completely described (without any ambiguity), and 250 puzzles with different level of ambiguity. Part of these puzzles are created by us, and part of them were collected from online sources, mentioned in the dataset.^{8 9}

3.2 Knights and Knaves Puzzles

The second part of our dataset is built on 300 puzzles with knights and knaves taken from <https://philosophy.hku.hk/think/logic/knights.php>. The complexity of these puzzles depends on the number of the individuals (ranging from 2 to 9) and on the type of sentences to be translated into FOL. Consider the Puzzle 3 with 4 inhabitants:

Puzzle 3 On the island where each inhabitant is either a knave or a knight, knights always tell the truth while knaves always lie. You meet four inhabitants: Bart, Dave, Rex and Sue. Bart tells you that Rex and Dave are both knights or both knaves. Dave says that Sue is a knave. Rex claims that Bart is a knave. Sue claims that Rex is a knight and Dave is a knave. Who is a knight and who is a knave?

Firstly, we formalise the common knowledge in all knights and knaves puzzles. The sentence *"The inhabitants are either knights or knaves"* can be translated into FOL with $\forall x (inhabitant(x) \rightarrow knight(x) \vee knave(x))$. The sentence *"One cannot be a knight and a knave in the same time"* is formalised with $\forall x (knight(x) \leftrightarrow \neg knave(x))$. A message $m(x)$ conveyed by a knight x is always true: $knight(x) \rightarrow m(x)$. Similarly, a message $m(x)$ conveyed by a knave x is always false: $knave(x) \rightarrow \neg m(x)$. This background knowledge appears in Listing 4.

⁸ The PuzzTe dataset is available on [Kaggle](https://www.kaggle.com).

⁹ The code for generating the PuzzTe dataset is available at <https://bitbucket.org/RoxanaSz/puzzte>.

Second, we automatically translate the puzzle into a FOL theory. We learn that there are four inhabitants, *Sue*, *Bart*, *Rex* and *Dave*: $\text{inhabitant}(\text{Sue}) \wedge \text{inhabitant}(\text{Bart}) \wedge \text{inhabitant}(\text{Rex}) \wedge \text{inhabitant}(\text{Dave})$. Hence, we compute a domain size of four. The message of *Sue* is: $m(\text{Sue}) \leftrightarrow \text{knight}(\text{Rex}) \wedge \text{knave}(\text{Dave})$, the message of *Dave* is: $m(\text{Dave}) \leftrightarrow \text{knave}(\text{Sue})$, the message of *Rex* is: $m(\text{Rex}) \leftrightarrow \text{knave}(\text{Bart})$, while the message of *Bart* is: $m(\text{Bart}) \leftrightarrow (\text{knight}(\text{Rex}) \wedge \text{knight}(\text{Dave})) \vee (\text{knave}(\text{Rex}) \wedge \text{knave}(\text{Dave}))$.

To obtain the above theory, we formalise the grammar rules depicted in Listing 3, that specify how words from different parts of speech are connected. We rely on the SEM feature, that defines the semantic of the words based on the lambda operator. For instance, $\lambda x. (\text{inhabitant}(x) \wedge \text{knight}(x))$ represents the elements from the domain that are both inhabitant and knight too. The rule for compound sentences (line 5) parses statements like "*Bart tells you that Rex and Dave are both knights or both knaves*". Given to Mace4, it computes the single model from Figure 4.

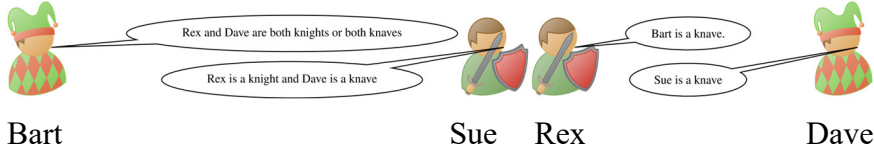


Figure 4 The single model computed by Mace4 for Puzzle 3

The resulted theory contains two unary predicates: $\text{knight}(x)$ and $\text{knave}(x)$ and four individuals: *Bart*, *Sue*, *Dave* and *Rex*. Hence, we generate $2 \times 4 = 8$ questions. Among them, 4 will generate entailments, and 4 will generate contradictions.

Table 2 Generating answers from ambiguous Knights and Knaves puzzles

Question	Theorem in FOL	Answer
Is Sue a knight ?	$\text{knight}(\text{Sue})$	Entailment
Is Dave a knave ?	$\text{knave}(\text{Dave})$	Entailment
Is Bart a knave ?	$\text{knave}(\text{Bart})$	Unknown
Is Rex a knight ?	$\text{knight}(\text{Rex})$	Unknown
Is Rex a knave ?	$\text{knave}(\text{Rex})$	Unknown
Is Bart a knight ?	$\text{knight}(\text{Bart})$	Unknown
Is Dave a knight ?	$\text{knight}(\text{Dave})$	Contradiction
Is Sue a knave ?	$\text{knave}(\text{Sue})$	Contradiction

A sample of these 8 questions are listed in Table 1. The generated formalisation in FOL appears in Listings 4 and 5. Note that the knowledge is divided into two

modules: the background knowledge for the knight and knaves puzzles, that is reused for all puzzles in the domain, and the specific information extracted from Puzzle 3.

```

S [SEM=<?np (?vp) >] ->
  NP [NUM=?n, SEM=?np] VP [NUM=?n, SEM=?vp]

S [SEM=?vp] ->
  Pron [NUM=?n] VP [NUM=?n2, SEM=?vp]

S [SEM=<?that (?sent1, ?sent2) >] ->
  S [SEM=?sent1]
  Conj [BOUND=sentences, TYPE=thatClause, SEM=?that]
  CompoundSent [SEM=?sent2]

CompoundSent [SEM=<?conjEither (?sent1, ?sent2) >] ->
  Conj [+either, SEM=?conjEither]
  S [SEM=?sent1]
  Conj [BOUND=sentences, SEM=?conj2]
  S [SEM=?sent2]

CompoundSent [SEM=<?conj2 (?sent1, ?sent2) >] ->
  Quantifier [+BOTH]
  S [SEM=?sent1]
  Conj [BOUND=sentences, SEM=?conj2]
  S [SEM=?sent2]

PP [TYPE=?type] -> P [+on] NP [TYPE=?type, NUM=?n]

NP [TYPE=where, NUM=?n] ->
  Det [SEM=?det] Nom [NUM=?n, SEM=?nom]

NP [NUM=?n, SEM=?nom] -> Nom [NUM=?n, SEM=?nom]

NP [LOC=?l, NUM=?n, SEM=?np] ->
  PropN [LOC=?l, NUM=?n, SEM=?np]

NP [NUM=?n, SEM=?nom] ->
  Numeral [NUM=?n] Nom [NUM=?n, SEM=?nom]

VP [NUM=?n, SEM=<?v (?prd) >, TYPE=one] ->
  AuxP [COP=?c, NUM=?n, SEM=?v, TYPE=one]
  Pred [SEM=?prd]

VP [NUM=?n, SEM=?vp] -> SV [NUM=?n, SEM=?vp]

VP [NUM=?n, SEM=?vp] -> Modal [] SV [NUM=?n, SEM=?vp]

VP [NUM=?n, SEM=?vp] -> Modal [] SV [NUM=?n, SEM=?vp] Pron [NUM=sg]

VP [NUM=?n, SEM=?vp] -> SV [NUM=?n, SEM=?vp] Pron [SEM=?pron]

```

Listing 3 Grammar rules for Puzzle 3

To extend the dataset with unknown questions, we generate ambiguous puzzles by removing a part of the information. For instance, by removing *Sue claims that Rex is a knight and Dave is a knave* from Puzzle 3, Mace 4 will generate two different models: in the first one, *Dave and Rex are knaves and Bart and Sue are knights*, and in the second model: *Rex and Sue are knights and Bart and Dave are knaves*. For this ambiguous puzzle, for 2 questions Mace4 generates entailments, for 2 questions Mace4 generates contradictions, and there are 4 questions that have unknown result (i.e., Mace4 generates models in which those statements are both true or false - see Table 2).

```
all x (inhabitant(x) -> knight(x) | knave(x)).
all x ((knight(x) -> -knave(x)) & (knave(x) -> -knight(x))).
knight(x) -> m(x).                               knave(x) -> -m(x).
```

Listing 4 Reusing knowledge for the knight and knaves puzzles

```
inhabitant(Sue) & inhabitant(Bart) & inhabitant(Rex) &
inhabitant(Dave).

m(Bart) <-> (knight(Rex) & knight(Dave)) | (knave(Rex) &
knave(Dave)).

m(Dave) <-> knave(Sue).

m(Rex) <-> knave(Bart).

m(Sue) <-> knight(Rex) & knave(Dave).

Sue!=Rex & Sue!=Bart & Sue!=Dave & Rex!=Bart & Rex!=Dave &
Bart!=Dave.
```

Listing 5 Generating the FOL theory for Puzzle 3

3.3 Zebra Puzzles

The third part of the PuzzTE dataset includes zebra puzzles, whose complexity depends on the number of the individuals or number of houses and on the type of hints/clues to be translated into FOL.

Puzzle 4 *There are 5 houses in five different colors. In each house lives a person with a different nationality. These five owners drink a certain type of beverage, smoke a certain brand of cigar and keep a certain pet. No owners have the same pet, smoke the same brand of cigar or drink the same beverage. The question is: Who owns the fish?*

Table 3 Question answering for Puzzle 4

Hypothesis			Theorem in FOL	Answer
German lives in house D.	<i>german(D)</i>	Entailment		
The house B is blue.	<i>blue(B)</i>	Entailment		
The man in house A drinks water.	<i>water(A)</i>	Entailment		
The man in house D owns the fish.	<i>fish(D)</i>	Entailment		
The house A is blue.	<i>blue(A)</i>	Contradiction		
Dane lives in house E.	<i>dane(E)</i>	Contradiction		
The man in house A drinks coffee.			<i>coffee(A)</i>	Contradiction
The man in house C owns the horse.			<i>horse(C)</i>	Contradiction

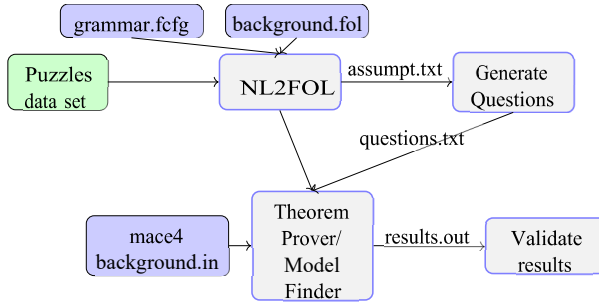


Figure 5 Generating the dataset from natural language puzzles

The clues of the puzzle appear in Listing 6. For instance, "The green house is on the left of the white house." is translated: $green(x) \& white(y) \rightarrow rightneighbor(y,x)$ (line 4 in Listing 6). we used auxiliary predicates like *differentFrom* or *rightneighbor*. To translate the sentence *In each house lives a person with a different nationality* we used the symmetrical relation $differentFrom(x,y) \rightarrow differentFrom(y,x)$, and we implement the unique name assumption with statements like $differentFrom(a,b)$ for all individuals in the domain. To define the relation between houses, we have: $rightneighbor(x,y) \vee rightneighbor(y,x) \leftrightarrow neighbor(x,y)$, which means that one is the neighbor of someone either if one lives just to his right or he lives just to your right. For "Each

house has at least one nationality, pet, drink, color, car.”, we have sentences like $\text{brit}(x) \wedge \text{swede}(x) \wedge \text{german}(x) \wedge \text{dane}(x) \wedge \text{norwegian}$.

```

brit(x) <-> red(x) .                               %the Brit lives in the red
house

swede(x) <-> dog(x) .                               %the Swede keeps dogs as pets
dane(x) <-> tea(x) .                                %the Dane drinks tea

green(x) & white(y) -> rightneighbor(y,x) .
green(x) <-> coffee(x) .
pallmall(x) <-> bird(x) .
yellow(x) <-> dunhill(x) .
milk(c) .
norwegian(a) .
blend(x) & cat(y) -> neighbor(x,y) .
horse(x) & dunhill(y) -> neighbor(x,y) .
bluemaster(x) <-> beer(x) .
german(x) <-> prince(x) .
norwegian(x) & blue(y) -> neighbor(x,y) .
blend(x) & water(y) -> neighbor(x,y) .

```

Listing 6 A FOL theory used to provide Yes/No answers

The resulted theory contains 25 unary predicates: 5 unary predicates for nationality, 5 for colors, 5 for pets, 5 for drinks, and 5 for cigars, five houses: *A, B, C, D, and E*. Hence, there are $5 \times 25 = 125$ questions. Among them 25 will generate entailments, and 100 will generate contradictions. A sample of these 8 questions are listed in Table 3. To extend the dataset, ambiguous puzzles are generated by removing clues. For instance, by removing the clue “*The man living in the center house drinks milk*” (line 8 from Listing 6), Mace4 generates 3 models. By removing “*The Norwegian lives in the first house*” (line 9), Mace4 generates 17 different models.¹⁰

¹⁰ Moreover, the set of zebra puzzles can be easily increased by adding new zebra puzzle that can be generated automatically, e.g., <http://new.mensus.net/brain/logic.shtml>.

3.4 System Overview

The architecture for building the PuzzTE dataset has four modules (Figure 5): (1) translating puzzles from natural language to FOL; (2) generating the entire set of questions for each puzzle; (3) computing answers by using theorem Mace4 model finder; (4) interpreting the results and generating the data set.

The first component takes each puzzle and translate line by line from natural language to FOL using the Natural Language Toolkit [Perkins, 2014]. The text is parsed using three grammars manually build for each puzzle domain: the fcfg grammar file (feature grammar file /context free grammar file), and a resource file with domain knowledge. Then, we generate the input file with assumptions for Mace4/Prover9 (e.g. Listings 4 or 5). The set of questions is generated automatically based on the assumptions in FOL, other existing predicates in the current puzzle.

Using the assumptions generated with the first module, and adding background knowledge file, we obtain the result for each question from questions list using Mace4/Prover9. The background knowledge for comparison puzzles was built based on the transitivity (see Listing 7), anti-symmetry, irreflexivity and trichotomy properties.

```
(shorter(x,y) & shorter(y,z) -> shorter(x,z)).           %transitivity

(taller(x,y) & taller(y,z) -> taller(x,z)).

exists x (tallest(x) & (all y (tallest(y) -> y=x))).

exists x (shortest(x) & (all y ((shortest(y) -> x=y)))).

all x all y (taller(x,y) & taller(y,x) -> x=y).           %antisymmetry

all x all y (shorter(x,y) & shorter(y,x) -> x=y).

-taller(x,x & -shorter(x,x).                             %irreflexivity

all x all y (shorter(x,y) | x=y | taller(x,y)).          %trichotomy

(shorter(x,y) -> (x!=y & -taller(x,y))).

(x=y) -> (-shorter(x,y) & -taller(x,y)).

(taller(x,y) -> (x!=y & -shorter(x,y))).

exists x (all y (-shorter(y,x)) & shortest(x)).

exists x (all y (-taller(y,x)) & tallest(x)).

taller(x,y) <-> shorter(y,x).
```

Listing 7 Adding background knowledge for comparison puzzles

Table 4 Quantifying the PuzzTE dataset

Puzzles	Complete information				Ambiguity (incomplete information)				
	No	Questions	Ent.	Contr.	No	Questions	Ent.	Contr.	Unknown
Comparison	50	3,056	1,130	1,926	250	15,680	4,353	8,389	2,938
Knights and knaves	300	430	215	215	150	940	105	105	730
Zebra	5	650	130	520	1	125	9	36	80
Total	355	4,136	1,475	2,661	401	16,745	4,467	8,530	3,748

Based on the result returned by Mace4 and Prover9 [McCune, 2005], we compute the answer for each question. When Mace4 generates a single model, in case of unambiguous puzzle, the current question is entailed from puzzle. For ambiguous puzzles, Mace4 can generate multiple solutions - in this case we have the unknown relation. The current version of the PuzzTE dataset contains 16,745 pairs from which 4,467 are labelled with entailment, 8,530 with contradiction, and 3,748 unknown (see Table 4).

4 Discussion and Related Work

By translating puzzles into FOL and then ask Mace4 for satisfiable models, our approach is also an automatic solution for solving puzzles. Solving logical puzzles is considered a challenging task, both for the human agent and the software agent. For the human agent, one can browse the 140 puzzles from the TPTP collection or the 144 puzzles modelled in FOL by Groza [2021]. For the software agent, several puzzle solvers have been proposed by Lev *et al.* [2004], Milicevic *et al.* [2012], Bogaerts *et al.* [2020], De Cat *et al.* [2018], Jabrayilzade and Tekir [2020], Mitra and Baral [2015], and Groza and Nitu [2021].

Lev *et al.* [2004] have proposed a solution based on grammar rules, FOL, and model builders. Lev *et al.* have focused on puzzles of type "multiple-choice question", so the inferences have to just find out the correct answer, not to discover it. Milicevic *et al.* [2012] have tackled the task by using the ink Grammar general-purpose English parser, a semantic translator, and an automated logical analyser. The solver is designed around the Zebra puzzles and tested on a dataset of 68 puzzles. Bogaerts *et al.* [2020] have proposed a solver for logic grid puzzles which also makes use of QA and XAI. De Cat *et al.* [2018] have developed the IDP system based on an extension of FOL. A drawback is that the named entities (e.g., persons, colours) cannot be detected automatically and must be stated by the user. Also, there exists a semi-automated process to detect the synonymy between verbs. Jabrayilzade and Tekir [2020] have proposed the DistilBERT tool that automatically solves logic grid puzzles. The clues are translated in Prolog. The

zebra puzzles have different categories that needs to be recognised within the text (e.g., person, name, occupation, colour). Mitra and Baral [2015] have developed the Logicia system also for 150 Zebra puzzles. The clues are classified using a maximum entropy model based on features like POS tags or dependency trees. The target language is answer set programming, based on which 71 out of 100 puzzles have been solved. Groza and Nitu [2021] have also used grammar rules and named entity recognition to obtain a theory in FOL, which was given to Prover9 theorem prover. The grammar rules were manually created by analysing 43 puzzles for identifying the recurrent predicates, and then tested the resulted grammar on 331 puzzles. NER has been used to identify the characters in the puzzle (e.g., number of knights and knaves). A graphical representation of the proof is displayed in order to explain why an answer to the puzzles is correct or wrong.

5 Conclusion

We propose here the PuzzTE dataset for textual entailment tasks. The three labels - entailment, contradiction and unknown - are automatically computed using model finders. This is the main distinguishing feature from the existing datasets, where labelling is performed by human experts or crowdsourcing. The PuzzTE dataset exploits two properties of puzzles: (i) each piece of information is necessary and (ii) no unnecessary information is provided. In our view, these properties make puzzles interesting candidates for machine comprehension tasks.

Ongoing work consists of organising a competition around this dataset on Kaggle. The competition will include two tasks: (1) questions answering with complete information, and (2) questions answering with incomplete information. From the available pairs, 80% will be available for training and validation, while the remaining 20% for testing the submitted models. We expect that machine learning approaches will face some difficulties, while approaches based on natural language understanding will be favoured by our inference oriented dataset. However, both machine learning and symbolic reasoning approaches will benefit from the fact that the puzzles do not contain unnecessary information.

The assure correct labels for each pair, the dataset was constructed in two steps. First, the puzzles were automatically translated from text to FOL via grammar rules. Second, each translation was manually verified to identify possible translation errors. Given, the FOL theory, we use Mace4 to correctly computes the entailment, contradiction or ambiguity relations between the initial text and the answer. We are currently working on building a question-answering dataset without need of human checking. For this, we can employ a puzzle generator that builds puzzles directly in FOL. Then, we verbalise the FOL theory into natural language. Hence, we will have both the FOL theory and the natural language puzzle.

References

- Bogaerts, B., Gamba, E., Guns, T. (2020). A framework for step-wise explaining how to solve constraint satisfaction problems. *arXiv preprint arXiv:2006.06343*.
- Bowman, S., Angeli, G., Potts, C., Manning, C. (2015). A large annotated corpus for learning natural language inference.
- Dagan, I., Glickman, O., Magnini, B. (2005). The pascal recognising textual entailment challenge. pages 177–190, 01.
- Cat, B., Bogaerts, B., Bruynooghe, M., Janssens, G., Denecker, M. (2018). Predicate logic as a modeling language: the idp system. In *Declarative Logic Programming: Theory, Systems, and Applications*, pages 279–323.
- Groza, A., Nitu, C. (2021). Natural language understanding for logical games. *arXiv preprint arXiv:2110.00558*.
- Groza, A. (2021). *Modelling Puzzles in First Order Logic*. Springer.
- Hickl, A., Bensley, J., Williams, J., Roberts, K., Rink, B., Shi, Y. (2005). Recognizing textual entailment with lcc’s groundhog system. In *In Proc. of the Second PASCAL Challenges Workshop*.
- Jabrayilzade, E., Tekir, S. (2020). Lgpsolver-solving logic grid puzzles automatically. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1118–1123.
- Khot, T., Sabharwal, A., Clark, P. (2018). Scitail: A textual entailment dataset from science question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr.
- Kotlerman, L., Dagan, I., Magnini, B., Bentivogli, L. (2015). Textual entailment graphs. *Natural Language Engineering*, -1:1–26, 06.
- Lev, I., MacCartney, B., Manning, C., Levy, R. (2004). Solving logic puzzles: From robust processing to precise semantics. In *Proceedings of the 2nd Workshop on Text Meaning and Interpretation*, pages 9–16.
- Magnini, B., Zanolli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T., Pado, S., Stern, A., Levy, O. (2014). The excitement open platform for textual inferences. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 43–48, Baltimore, Maryland, June. Association for Computational Linguistics.
- Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, pages 216–223, Reykjavik, Iceland, May. European Language Resources Association (ELRA).
- McCune, W. (2005). Prover. University of New Mexico.
- Milicevic, A., Near, J., Singh, R. (2012). Puzzler: An automated logic puzzle solver. *Dostopno na: <http://people.csail.mit.edu/jnear/puzzler/writeup.html>*.
- Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J. (2021). Deep learning based text classification: A comprehensive review.
- Mitra, A., Baral, C. (2015). Learning to automatically solve logic grid puzzles. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1023–1033.

- Perkins, J. (2014). Python 3 text processing with NLTK 3 cookbook. Packt Publishing Ltd.
- Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification.
- Wagner, W. (2010). Steven bird, ewan klein and edward loper: Natural language processing with python, analyzing text with the natural language toolkit. *Language Resources and Evaluation*, 44:421–424.
- Williams, A., Nangia, N., Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana, June. Association for Computational Linguistics.
- Yang, Y., Yih, W., Meek, C. (2015). WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, September. Association for Computational Linguistics.

Building a Corpus of Medical Paraphrases in Romanian

IOANA BUHNILĂ

*LiLPa UR 1339 – Linguistique, Langues, Parole
University of Strasbourg, France*

`ioana.buhnila@etu.unistra.fr`

Abstract

Understanding medical language is a challenge for laypeople due to complex scientific formula and technical terms used in this domain. These specific terms are difficult to understand and require paraphrasing (rewriting, simplifying) for laypeople. In this preliminary study, we aim to constitute a first medical corpus, annotated with medical paraphrases in Romanian. This resource is useful for the simplification of monolexical ("hipoacuzia"; hypoacusis) and multiword medical terms in Romanian ("carcinom endometrial"; endometrial carcinoma). Our method consists of automatically searching medical terms and specific markers that help to paraphrase them. We present our methodology to collect this corpus and for the annotation process. We also present quantitative data and first experiences on qualitative annotation of the identified paraphrases.

Key words - medical paraphrasing in Romanian, simple and multiword medical terms, paraphrase markers, medical paraphrase corpus.

1 Introduction

Medical terms are difficult to understand for laypeople. The word *term* represents a lexical unit that expresses concepts specific to a field of knowledge, recognised and shared by members of a community of specialists [Costa, 2005]. To understand technical terms, one needs to be familiar with the domain they belong to. Understanding the language system in which they are written is not sufficient [Coşeriu, 1967]. The *term* belongs to an autonomous “subsystem” of the language and its goal is to communicate technical or scientific knowledge [Contente, 2005].

The opacity of medical terms is given by their Greek and / or Latin etymology. They are often composed of a mix of prefixes / suffixes from these two ancient languages applied to modern words. We clearly observe the difficulty that laypeople have to understand medical terms such as “cholecystectomy”: which is formed with two Greek basis, “chole” (= bile) and “ectomy” (= surgical removal), and in the middle of these, there is a Latin basis, “cystis” (= bladder) [Grabar and Hamon, 2015]. Medical terms might be simplified by using synonyms from the common language, but it is sometimes difficult to find the right synonyms for

them. Therefore, explanation and paraphrasing are required to ease the understanding of medical texts for the laypeople. *Paraphrasing* is the process of rewriting in order to explain or simplify a sentence or phrase.

In this paper, we assume that medical texts written for laypeople, published on dedicated websites for non-specialists, contain a large number of paraphrases. These types of articles aim to explain scientific medical concepts (diseases, treatments, medical procedures) using simpler words and expressions from the common language that lay readers can understand.

In this study, we aim to constitute a first medical corpus annotated with medical paraphrases in Romanian. This resource is useful for the simplification of monolexical (one word) and multiword medical terms in Romanian (two words and above / expressions). Our method consists of automatically searching medical terms and specific markers that help to paraphrase them, like in the example: “Grăsimile saturate și trans pot crește nivelul <term>colesterolului (LDL)</term> (<marker>cunoscut sub numele de</marker> <paraphrase>colesterol rău</paraphrase>) în sânge.”¹ (Saturated and trans fats can increase the level of cholesterol (LDL) (known as bad cholesterol) in the blood.) Section 2 presents related work on paraphrase identification in the medical domain as well as on paraphrase markers. In Section 3 we present our methodology for the creation of the medical paraphrase corpus and the extraction of the medical terminology, while Section 4 shows our first results. We conclude with the potential use of our corpus and further qualitative analysis work.

2 Related Work

To our knowledge, there is no corpus of medical paraphrases available for Romanian, besides a manually collected corpus of scientific and lay medical texts of 42,140 tokens to identify scientific terms and their simpler synonyms in lay texts [Buhnila, 2018]. This corpus has been used to create a glossary of 113 medical terms and their common language synonyms. The term search was performed semi-automatically with the help of a reference corpus for the common language and n-gram patterns (multiword expressions).

2.1 Medical Resources in Romanian

There are several resources for the medical domain in Romanian such as the biomedical corpus BioRo [Mitrofan and Tufiş, 2018] and the gold standard medical corpus MoNERo [Mitrofan *et al.*, 2019] which is morphologically annotated with part of speech (PoS) tags and with entities specific to the medical domain. The SiMoNERo [Barbu Mititelu and Mitrofan, 2020] medical treebank was created on top of the work on the MoNERo corpus by adding the syntactic level of annotation. In our work we use the medical entities annotated in MoNERo

¹ Original text from our corpus has no Romanian diacritics.

[Mitrofan *et al.*, 2019] to extract a list of simple and multiword medical terms in Romanian.

2.2 *Work on Medical Paraphrases and Term Definition*

Several research projects study medical paraphrases in other languages. We consider previous work on French, a Romance language whose medical terminology is close to the Romanian one. Grabar and Hamon [2016] searched for medical terms provided by existing medical terminologies (Snomed International [Côté, 1996] and the French part of UMLS (Unified Medical Language System)² [Lindberg *et al.*, 1993]) in a corpus of Wikipedia articles. Their study focuses on paraphrases that appear in free contexts, without being in close co-occurrence with the medical term being searched (technical terms and paraphrases can be separated by several words). Grabar and Hamon [2016] use the French morphological analyser Dérif [Namer, 2009] to extract the actual words in modern French from medical terms of Greek or Latin origin. For example, the term “myocardique” contains the modern French words “muscle” / muscle (myo) and “cœur” / heart (carde). These words were projected onto the corpora and their defining contexts were automatically extracted using patterns such as “is”, “defined as” [Péry-Woodley and Rebeyrolle, 1998]. After a manual evaluation of the quality of the paraphrases, Grabar and Hamon [2016] find paraphrase equivalents, definitions, and explanations in general language for 2,596 unique medical terms.

Săpoiou [2013] studied the lexical relations of hyponymy and hypernymy in the definitions of medical terms as they appear in general Romanian language dictionaries. In these definitions, the term is the hyponym (a specific word), while the definition starts with or contains a hypernym (the generic word) [Bidu-Vrâncanu, 2007]. Săpoiou [2010] proposed a multi-level classification of hypernyms in the definition of a medical term, starting from the most general (“cardiomalacie: distrofie” / cardiomalacia: dystrophy), to the most specific difference: determiner (“generalizată” / generalized), anatomy (“a miocardului” / of the miocardium), process / consequence / cause (“prin infiltrarea fibrei musculare cu grăsimi” / by infiltrating muscle fiber with fat).

Barbu Mititelu [2011] worked on the lexical relation of hyponymy / hypernymy applied on journalistic and medical texts³ by automatically identifying two types of patterns: (1) “HYPERNYM hyponymy pattern HYPONYM” and (2) “HYPONYM hyponymy pattern HYPERNYM”. The best results in terms of accuracy (more than 66,66%) in the medical domain were obtained with the following patterns: GN (group noun), adică / “that is” (marker) GN (group noun); GN “care fi” / that be GN; GN “în special” / especially GN; GN “inclusiv” /

² UMLS is a set of files and software that brings together health and biomedical vocabularies for automatic use on different computer systems (<https://uts.nlm.nih.gov/uts/>).

³ EMEA (European Medicines Agency) corpus of 11,914,802 tokens.

inclusively GN; GN “precum” / as GN; GN “(în) afară de” / besides GN; GN “de obicei” / usually GN; GN “ci (și/doar)” / but (also/only) GN.

While Grabar and Hamon [2016] focus on simple neoclassical compounds (such as “cholecystectomy”), our method does not include work on the neoclassical compounds. We search simple and multiword medical terms extracted from the MoNERo corpus [Mitrofan *et al.*, 2019] in our own corpus. Our method is also different from Buhnilă’s [2018], as we use a list of medical terms for term identification, and also, we look for complex paraphrases instead of common language synonyms. The work of Săpoiou [2013] and Barbu Mititelu [2011] constitutes an important reference for our process of lexical annotation of the medical paraphrases.

3 Methodology

We used a list of simple and multiword medical terms extracted from the MoNERo annotated corpus [Mitrofan *et al.*, 2019] to automatically identify them in our corpus using our own programming scripts in Perl code. Then, paraphrase markers are automatically identified in the terms’ context. Paraphrase markers for Romanian are inspired from the literature [Péry-Woodley and Rebeyrolle, 1998; Vassiliadou, 2016; Eshkol-Taravella and Grabar, 2018], from translations to Romanian and from our own observations on our corpus. Several tags (<term>; <marker>; <paraphrase>) were implemented for the automatic data analysis. We proceed to a detailed analysis of the data, to identify the lexical relations (synonymy, hypernymy, hyponymy, meronymy) and the semantic-pragmatic relations (definition, designation, exemplification, explanation, rephrasing). We assume that some markers are correlated with specific relations. A first hypothesis about the generic paraphrase marker “este o/este un” (“is a”) is tested. We assume that it introduces medical paraphrases with the lexical relation of hypernymy and the semantic-pragmatic relation of definition. The application of our methodology is presented in detail as it follows.

3.1 Medical Corpus

Our corpus is collected through the web crawling tool Sketch Engine [Kilgarrieff *et al.*, 2014]. We created a corpus of 990 thousand words (73 thousand sentences) from the medical website *sfaturimedicale.ro* that contains medical articles written for laypeople. The collected corpus is cleaned from the tags imported by default by the tool (<p>; </p>).

3.2 Medical Term List

We created a medical terminology from the MoNERo annotated corpus [Mitrofan *et al.*, 2019]. The MoNERo corpus is annotated with parts of speech (PoS) and named entities (NE) from the medical domain. The corpus contains texts from

three medical subdomains: cardiology, diabetes, endocrinology. The texts are extracted from scientific books, journal articles and blog posts. MoNERo contains 154,825 tokens (all morphologically tagged), where 14,133 medical concepts are annotated with four UMLS semantic categories [Bodenreider, 2004]: (i) 1,964 for body parts (ANAT); (ii) 4,156 for chemicals and drugs (CHEM); (iii) 6,611 for disorders (DISO); (iv) 1,402 for procedures (PROC).

We used the codes for the semantic categories that mark the beginning (B-DISO) and the intermediary/end of the medical term (I-DISO) to extract a medical terminology. We then deleted the duplicates in order to have only unique terms. Out of the 14,133 medical terms extracted, we obtained 7,528 unique values.

We eliminated from the list several types of elements:

- *the word identified as the term "a"* (B-DISO; Va--3s) (as form of the verb "a avea", *to have*) and the verb "este" (*to be*, B-CHEM; Vaip3s), which distort searches and tagging of terms;
- *orthographic elements* that distort the results, as they are not medical terms (/ B-CHEM SLASH; . B-ANAT PERIOD; - B-CHEM DASH; , B-CHEM COMMA);
- *foreign language terms and abbreviations*, mostly English (X class), such as: speckle; tibialis; glutation;
- *abbreviations* (Y, Ya, Yc, Yn, Rc, Mc classes), such as: AIRglucose; FSIVGTT; TNG; AG; Ser, etc. For an initial tagging, we have discarded abbreviations, as abbreviations such as *K* for potassium, *Ca* for calcium create false term tags. We will do another automatic search in a future work to identify abbreviations.
- *functional words* (Spsa class): din, de (from), cu (with).

After all these words were deleted from the list, we obtained a medical terminology in Romanian of 6,899 simple and multiword medical terms. We analysed this list and we found that some multiword terms are identified as such, but they are in fact simple terms followed by functional words (prepositions and conjunctions) or even by punctuation signs: "atrofie a"; "ventriculare și"; "proinsulinei și"; "pancreasului."; "ruptura de"; "rupture de"; "sindroamelor de"; "deficite de". Therefore, we counted these terms as simple medical terms.

Table 1 Medical term list extracted from MoNERo

Medical terms in MoNERo	Medical terms extracted	Unique medical terms	Medical terms after analysis	Simple medical terms	Multiword medical terms
14,133	14,133	7,528	6,899	2,434	4,465

3.3 Paraphrase Markers

We performed an automatic search for typical paraphrase markers in Romanian, by applying the methods used for French language [Péry-Woodley and Rebeyrolle, 1998; Chéria, 2010; Vassiliadou, 2016; Eshkol-Taravella and Grabar, 2018; Fuchs, 2020]. French and Romanian are Romance languages, and they share the same etymology of medical terms. Therefore, we searched for the Romanian translation of paraphrase markers identified in French, such as “autrement dit” (*cu alte cuvinte*), “signifie” (*înseamnă*), “est une maladie” (*este o boală*). We found other paraphrase markers through our own observations on the Romanian corpus: “sub denumirea de / sub numele de” (*under the name of*), “este cunoscut” (*is known*), “cunoscut și” (*also known as*), “cum ar fi” (*such as*), “mai precis” (*more precisely*), “este termenul” (*is the term*). We looked for both forms of the word, with and without diacritics (*este o afecțiune / afectiune*), as texts on medical popularization websites and articles on the internet are often written without Romanian diacritics (ă, â, î, ș, ț).

3.4 Automatic Annotation Process

Our method consists of an automatic term identification with our own programming scripts. We used the medical terminology extracted from MoNERo and we searched this list of terms in our corpus. We then conducted a second automatic search of our list of paraphrase markers only within the sentences that contain at least one medical term.

3.5 Levels of Annotation

The texts are split one sentence per line considering phrase end orthographic markers (“.”, “!”, “?”) to facilitate annotation. While we did the automatic searches, we also inserted specific tags to create an annotated version of the medical paraphrase corpus. This version is used to extract the information for simplification purpose and for machine learning experiences. We create an annotation guide aimed to identify correct medical paraphrases.

Table 2 Romanian paraphrase markers

Romanian paraphrase markers	Translation
înseamnă / înseamna	means
mai exact	more precisely
mai precis	more precisely
cum ar fi	such as
cu alte cuvinte	in other words
este cunoscut/ cunoscută	is known
cunoscut/ cunoscută sub numele de	known as
cunoscut/ cunoscuta si / cunoscută și	also known as
sub denumirea de	under the name of
définit/ă, definiție/ definiție	defined, definition
este termenul	is the term
este o (boală/ boala)	is a disease
este o (afecțiune/ afectiune)	is a condition

The annotation scheme has the following levels:

- i) *Tagging medical terms* - we searched for terms formed by flexion (“pericardita constrictivă”/ *constrictive pericarditis*) and not the terms’ lemmas (“pericardită constrictiv / *constrictive pericarditis*) as our objective was to find multiword terms in discourse. In future experiments we will also search for lemmas (for simple and multiword terms) in our corpus and match them to the MoNERo term list in both flexion and lemma form. We marked with pair tags the medical terms identified from our list of terms from MoNERo. This tagging allows us to identify whether the terms annotated are paraphrased in the texts. In our manual sentence analysis, we will tag all the medical terms that are paraphrased (but not all medical terms, as the goal of our study is to identify paraphrased medical terms).
- ii) *Tagging paraphrase markers* – the annotation guide contains the list of possible paraphrase markers presented in subsection 3.3. These markers are also tagged in order to do several automatic analysis on our corpus.
- iii) *Annotations of medical paraphrases* – we want to annotate the type of paraphrase (simple noun, noun phrase, verbal phrase, a subordinate sentence, or a full sentence). We will also analyse these paraphrases from a lexical and semantico-pragmatic point of view:

- *lexical*: shows the lexical relation between the paraphrase and the medical term (synonymy, hyponymy, hypernymy, meronymy).
- *semantic-pragmatic*: expresses the reasons that motivate the writer to use paraphrases (definition, designation, exemplification, explanation, rephrasing). These definitions of semantic-pragmatic relations are inspired by Eshkol-Taravella and Grabar's [2017] taxonomy on oral texts of common language and adapted by us to written texts in the medical domain, as it follows:

- *definition*: the term is given a definition because it is considered to be too technical or domain specialised, thus difficult to understand (“<term>Tiroidita autoimună Hashimoto</term> <marker>este o <paraphrase>afecțiune</marker> în care sistemul imunitar atacă celulele tiroidei </paraphrase>.” / Hashimoto's autoimmune thyroiditis is a condition in which the immune system attacks thyroid cells.)

- *designation*: the term is paraphrased using another word or term (“<term>Oreionul </term> <paraphrase>(parotidita epidemică)</paraphrase> / Mumps (epidemic parotitis); where the brackets function as markers. We will consider orthographic markers in future analysis.

- *exemplification*: the paraphrase is a list of examples (several entities of the same type) that help to illustrate the meaning of the term (“<term>organele pelvine feminine</term>: <paraphrase>vaginul, colul uterin, pereții uterini, trompele uterine, ovarele sau țesuturile care înconjoară uterul, vezica urinară sau chiar intestinale</paraphrase>” / female pelvic organs: vagina, cervix, uterine walls, fallopian tubes, ovaries or tissues surrounding the uterus, bladder or even intestines).

- *explanation*: the term is explained through a particular situation or procedure (“<term>Hipotiroidismul</term> <marker>înseamna că<marker> <paraphrase>glanda tiroidă nu poate produce cantitatea normală de hormon tiroidian</paraphrase>.” / Hypothyroidism means that the thyroid gland cannot produce the normal amount of thyroid hormone),

- *rephrasing*: the meaning of the term is expressed with simpler words (“<paraphrase>Glanda leneșă</paraphrase> (<term>hipotiroidismul<term>)” / Sluggish gland (hypothyroidism).)

In this work we analysed only the lexical relation of hypernymy, and the semantico-pragmatic relation of definition. Future work will include analysis of all relations.

4 Results Analysis

After the automatic extraction, we obtained 3,067 sentences that contain 5,426 instances of medical terms and 3,144 instances of paraphrase markers. We analysed all occurrences of paraphrase markers in terms of number of instances found in the corpus and in term of relative frequency (the frequency divided by total number of instances). We present this analysis in Table 3. In our counting process, we considered markers that overlap with another, such as “sub denumirea de” and “cunoscut/ă/a sub denumirea de” being part of the more complex marker “este cunoscut/ă/a sub denumirea de”. We counted the longest marker each time and we did not count the shorter markers that overlap with the longer one (for example, “este cunoscut/ă/a” is not counted when it is part of the longer marker “este cunoscut/ă/a sub numele de”).

We then classified them from the most frequent, “este o (boală)” (is a disease), “cum ar fi” (such as), to the least frequent: “este cunoscut sub denumirea de” (is known under the name of). We precise that for the most frequent marker, “este o (boală) / is a disease” of 1,484 instances in the corpus, 1,325 of them are for the marker “(o) boala/ă / a disease” (and only 159 for the complete form of the marker “este o boala/ă / is a disease”). We considered this word as a marker because it helped us to identify medical paraphrases such as: “<term>sindrom Guillain-Barré</term>, <paraphrase><marker>o boală</marker> severă paralizantă</paraphrase>”, where “o boală / a disease” appears alone in the structure of an apposition, thus without the generic marker “este o” (is a). We also noticed that the marker becomes a required part in the medical paraphrase structure.

All the 3,067 sentences will be manually analysed in order to identify correct paraphrases. In this work we tested our first research hypothesis that concerns the role of paraphrase markers: all sentences with the generic paraphrase marker “este o/este un” (“is a”) are paraphrases with:

- i) *the lexical relation of hypernymy*, which means that the medical paraphrase is the hypernym of the medical term (the hyponym).

- ii) *the semantico-pragmatic relation of definition*, which means that the writer's intention and the purpose of the text is to give a simpler definition to the scientific medical term.

Table 3 Romanian paraphrase markers: instances and frequency among all identified markers

Romanian paraphrase markers	Translation	Number of instances found	Total of instances	Relative frequency
înseamnă / înseamna	means	0 / 169	169	0.05375
mai exact	more precisely	24	24	0.00763
mai precis	more precisely	7	7	0.00222
cum ar fi	such as	1,174	1,174	0.37340
cu alte cuvinte	in other words	6	6	0.00190
este cunoscut/ă/a	is known	36 / 0 / 11 (-8)	39	0.01240
cunoscut/ă/a și / cunoscut/ă/a si	also known as	0 / 0 / 0 / 8 / 0 / 27	35	0.01113
cunoscut/ă/a sub numele de	known as	17 / 0 / 15	32	0.01017
este cunoscut/ă/a sub numele de	is known under the name of	1 / 0 / 4	5	0.00159
sub denumirea de	under the name of	41 (-8)	33	0.01049
cunoscut/ă/a sub denumirea de	known under of the name of	0 / 0 / 8 (-3)	5	0.00159
este cunoscut/ă/a sub denumirea de	is known under of the name of	0 / 0 / 3	3	0.00095
définit/ă/a, definiție/tie	defined, definition	14 / 0 / 0, 0 / 5	19	0.00604
este termenul	is the term	3	3	0.00095
este o (boală/ boala)	is a disease	1 / 158 (2 / 1,484) (- 159)	1,325	0.42143
este o (afecțiune/ afectiune)	is a condition	0 / 156 (0 / 413) (- 156)	257	0.08174
Total			3,144	100.00%

To test the validity of this hypothesis, we annotated all the generic markers “este o /una; este un” (is a) in our corpus. In our searches for paraphrase markers, we used the forms “este o afecțiune” (is a condition), “este o boală” (is a disease), but also the medical paraphrase indicators “afecțiune” (condition), “boală”

(disease) separately. We decided not to look for the generic marker “este o /una; este un” (is a) alone in order to not have too many wrong results.

We found 365 occurrences of the marker “este o”, 6 for “este una” (feminine forms of the determiner), and 48 for “este un”, the masculine form of the determiner. These markers appeared in 423 sentences. We then analysed and annotated each sentence to see if these specific markers help identifying paraphrases with the semantic-pragmatic role of definition. Our results showed that 61 sentences were not medical paraphrases, as the sentence was incomplete: “<marker>Este o boală</marker> care afectează vederea centrală.” (Is a disease that affects central vision). As we can observe in this sentence, the medical term that is paraphrased does not appear in the same sentence, thus we cannot consider it a paraphrase. Another similar example is “Aceasta <marker>este o</marker> afecțiune caracterizată de lipsa celulelor nervoase din intestinul gros.” (This is a condition characterized by the lack of nerve cells in the large intestine). For this study, we concluded that sentences starting with “Este o/un/una” (Is a) or “Aceasta este” (This is) are not paraphrases, as we cannot identify the medical term that is paraphrased. Further work can include analysis of coreference chains in order to find the referent (medical term) of the anaphora “aceasta” (this) in previous contexts.

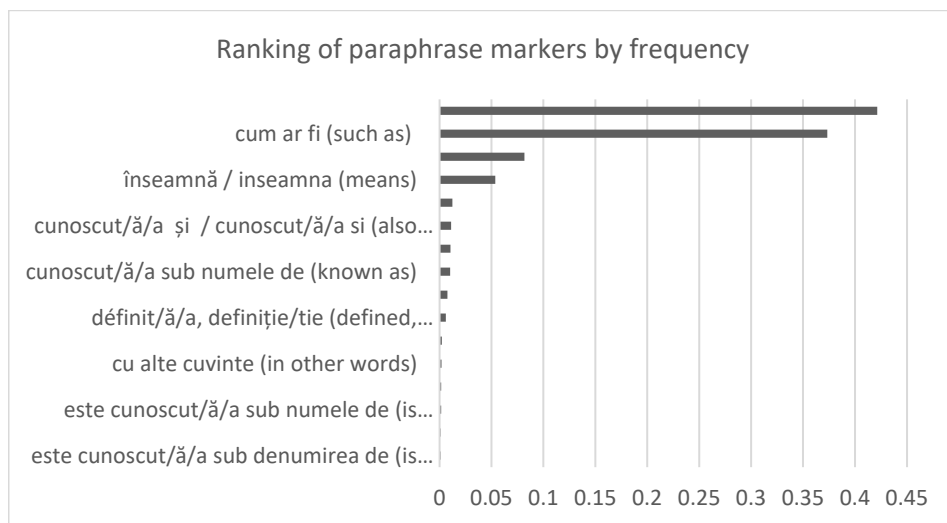


Figure 1 Romanian paraphrase markers classified by top frequency

Out of the manual analysis of the 423 sentences with the marker “este o/una/un” (is a), we obtained 362 sentences containing correct paraphrases that confirm our hypothesis.

5 *Access and Potential Use of the Corpus*

Preliminary work is still in progress as we want to conduct an in-depth analysis of all the sentences that can contain correct paraphrases. We will also extend our corpus with articles extracted from other scientific and medical popularization websites in Romanian using the tool Sketch Engine. Our goal is to have a large number of correct paraphrases (several thousands). A larger corpus could be used for natural language processing tasks (automatic medical term simplification, automatic and deep learning tasks). The final corpus will be available both in raw and annotated form, with a license for free use and development for researchers.

6 *Conclusions*

This work aims at creating a first medical paraphrase corpus in Romanian. Paraphrases are annotated in their context (full phrase) in order to allow further research and analysis. At this stage of our research, we have identified 3,067 phrases containing medical terms and paraphrase markers. Our future analysis of these sentences will enable us to create the first lexical resource of this kind for Romanian, being useful for research in medical text simplification.

For the moment, we analysed 423 sentences and identified 362 correct medical paraphrases with the lexical relation of hypernymy and semantic-pragmatic relation of definition, according to the theory of Eshkol-Taravella and Grabar [2017]. In further work we will include patterns identified by Barbu Mititelu [2011] to search for paraphrase markers such as “adică” (that is), “precum” (as). We are continuing this work of annotation as to find different types of medical paraphrases (explanation, designation, exemplification, explanation, rephrasing) with different lexical relations (synonymy, meronymy).

Our corpus could be useful for experiences and research in natural language processing (NLP) tasks such as texts simplifications, question answering, text generation. The annotated version of our work can be used for developing tools designed for patients and laypeople to better understand medical concepts (through chatbots for patients or medical popularization documentation). Future work includes experiments with machine learning techniques such as neural networks and transformers (Multilingual BERT [Devlin *et al.*, 2019]) in order to automatically identify medical paraphrases in Romanian.

Acknowledgements

This work is supported by a scholarship from the French Government granted to the French-Romanian doctoral project co-directed between the Research Unit (UR

1339) LiLPa (Linguistique, Langues, Parole) from the University of Strasbourg and the Research Institute of Artificial Intelligence of the Romanian Academy. The author thanks the researchers from LiLPa who helped improve this paper with their expert scientific advice and review.

References

- Barbu Mititelu, V. (2011). Hyponymy Patterns in Romanian. *Memoirs of the Scientific Sections of the Romanian Academy*, vol. XXXIV, pp 1-13.
- Barbu Mititelu, V., Mitrofan, M. (2020). The Romanian medical treebank-SiMoNERo. In *Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2020*, pp 7–16.
- Bidu-Vrâncanu, A. (2007). Lexicul specializat în mișcare. De la dicționare la texte. București. Editura Universității din București.
- Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research* 32 (90001), pages 267D - 270, <https://doi.org/10.1093/nar/gkh061>.
- Buhnila, I. (2018). Simplification lexicale entre les textes scientifiques et les textes de vulgarisation du domaine de la médecine. *Mémoire de Master*, Université de Strasbourg, 67000 Strasbourg, France.
- Chéria, N. (2010). Reformulation paraphrastique et non paraphrastique dans La Jalousie de Robbe-Grillet : l'exemple de c'est-à-dire vs en fait et en réalité. *L'Information grammaticale* 127, pp 43-47.
- Contente, M. (2005). Termes et textes : la construction du sens dans la terminologie médicale. In *Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*, pp 453-65, Bruxelles, Belgique.
- Coseriu, E. (1967). Structures lexicales et enseignement du vocabulaire. *Les Théories linguistiques et leurs applications*, Strasbourg, Conseil de l'Europe, pp9-51.
- Costa, R. (2005). Texte, terme et contexte. In *Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*, pp 79-88, Bruxelles, Belgique.
- Cote, R. (1996). Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. Université de Sherbrooke, Sherbrooke, Québec.
- Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*. <http://arxiv.org/abs/1810.04805>.
- Eshkol-Taravella, I., Grabar, N. (2017). Taxinomie dans les paraphrases du point de vue de la linguistique de corpus, *Syntaxe et Sémantique*, vol. 18, no. 1, pages 149-184.
- Eshkol-Taravella, I., Grabar, N. (2018). Paraphrases : de l'étude outillée dans les corpus disponibles vers leur détection automatique. *Langages* N° 212 (4), pages 5-16.
- Fuchs, C. (1982). La Paraphrase, *PUF*, Paris.
- Fuchs, C. (2020). Paraphrase et reformulation : un chassé-croisé entre deux notions. *Autour de la reformulation*, 36, Droz, pages 41-55, Coll. Recherches et Rencontres, 978-2-600-06051-6.

- Grabar, N., Hamon, T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. In *22ème Traitement Automatique des Langues Naturelles*, 14. Caen, France.
- Grabar, N., Hamon, T. (2016). Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. *Traitement Automatique des Langues*, Varia, 57, vol. 1, pp 85-109.
- Kilgarrieff, A. Vít Baisa, J., Miloš, J., Vojtěch, K., Michelfeit, J., Rychlý, P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, vol. 1, pp 7-36.
- Lindberg Donald, Humphreys Betsy and Mccray Alexa. 1993. *The Unified Medical Language System*, Methods Inf Med, vol. 32, no 4, pp 281-291.
- Mitrofan, M., Barbu Mititelu, V., Mitrofan, G. (2019). MoNERo: A Biomedical Gold Standard Corpus for the Romanian Language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp 71-79. Florence, Italy: Association for Computational Linguistics. <https://doi.org/10.18653/v1/W19-5008>.
- Mitrofan, M., Tufiş, D. (2018). BioRo: The biomedical corpus for the Romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, pp 1192-1196.
- Mitrofan, M., Barbu Mititelu, V., Mitrofan, G. (2019). MoNERo: A Biomedical Gold Standard Corpus for the Romanian Language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, AC. pp 71-79.
- Namer, F. (2009). Morphologie, Lexique et TAL : l'analyseur DériF. *TIC et Sciences cognitives*, Hermes Sciences Publishing, London.
- Péry-Woodley, M., Rebeyrolle, J. (1998). Domain and genre in sublanguage text: definitional microtexts in three corpora, *LREC*, pp 987-992.
- Săpoiou, C. (2010). Tipuri de hiperonime în definițiile lexicografice ale termenilor medicali. Edited by Rodica Zafiu, Adina Dragomirescu and Alexandru Nicolae. *Limba română: controversă, delimitări, noi ipoteze. Actele celui de-al 9-lea Colocviu al Catedrei de Limba Română*, I, Section Lexic, semantică, terminologii. Editura Universității din București, pp 297.
- Săpoiou, C. (2013). Hiponimia în terminologia medicală. Modalități de abordare în semantică și lexicografie, Pitești, Editura Trend, pp 199.
- Tufiş, D., Barbu Mititelu, V., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian. In *RRL*, LXIV, 3, pp 227-240.
- Vassiliadou, H. (2016). Mouvements de réflexion sur le dire et le dit: c'est-à-dire, autrement dit, ça veut dire. *Histoires de dire. Petit glossaire des marqueurs formés sur le verbe dire*, L. Rouanne & J.-C. Anscombe (éds), Bern/Berlin/Bruxelles/New York/Oxford/Wien, Peter Lang, pp 339-364.

CHAPTER 4

SPEECH AND TEXT MODELS FOR ROMANIAN LANGUAGE

A New Method of Emphasizing the Formants of the Vocal Spectrum

MARIUS ZBANCIOC^{1,2} and MONICA FERARU¹

¹*Institute of Computer Science, Romanian Academy, Iași*

²*Technical University “Gheorghe Asachi” of Iași*

zmarius@etc.tuiasi.ro,

monica.feraru@iit.academiaromana-is.ro

Abstract

In this paper, we propose a new method for emphasizing the spectral magnitude of the fundamental frequency F0 and formants F1-F4 from the vocal signal. MFSC (Mel-Frequency Spectral Coefficient) images with enhanced traces for F0, F1,..., F4, will be generated in order to improve the accuracy recognition of emotions when it is using deep-learning neural networks DL-CNN. The proposed algorithm can be generalized for any type of image (not necessarily spectrograms) allowing the highlighting of some areas of interest. We have made some simulation and the accuracy of emotion recognition rate was around 86% for emphasized MFSC images, which is better than those for standard MFSC images.

1 Introduction

Detecting formants is a complicated task. Different methods of detecting the fundamental frequency F0 and the formants F1...F4 are still work in the progress in order to obtain more robust and precisely algorithms. These algorithms can be used in voice activity detection VAD with applications in various domains (mobile communication services [Freeman, 1989], real-time speech transmission on the Internet [Karray and Martin, 2003], noise reduction for digital hearing aid devices, in speech processing: coding, enhancement and recognition. VAD can be used for an automatic speech recognition system [Javier Ramirez *et al.*, 2004]. During the last decade many researchers have studied different strategies for detecting speech in noise and the influence of the VAD decision on speech processing systems [Karray and Martin, 2003; Sangwan *et al.*, 2002; Freeman, 1989; Sohn, 1998].

Emotion recognition through speech has many potential applications; however, the challenge comes from achieving a high rate of emotion recognition while using limited resources or input signals with a high level of noise. Voice activity detection (VAD) is an important method for pre-processing speech data. A VAD works to detect the presence or the absence of human voice in a signal [Ong and Tan, 2016]. Research on practical applications of VADs and how to implement them is numerous, such as the paper published by [Sehgal and Kehtarnavaz, 2018]

proposing the implementation of real-time VAD Smartphone application or the algorithm proposed by [Song *et al.*, 2018] for assisting the hearing impaired. Research on person identification has been improved using VAD [Gokul and Lalitha, 2018].

In the paper [Alghifari *et al.*, 2019] the possibility of improving speech emotion recognition using the concept of voice activity detection (VAD) is explored. Researchers found that VAD improved the recognition rate of 5 emotions (happiness, anger, sadness, fear and neutral tone) by 3.7%, when recognizing clean signals, while the effect of using VAD when training a network with both, clean and noisy signals lead to an improvement of the previous results by 50%.

VAD can be applied to smart homes to automate everyday processes which are using voice commands. From such approaches can benefit a wide range of people, especially people who face challenges in performing everyday tasks such as operating electronic appliances in a home. This can include people living with physical disabilities or special needs, as well as the elderly; anyone would appreciate the convenience of automated processes in a home. People with special needs and older people, especially those living alone, have difficulty controlling appliances and often need another person to help them with everyday tasks such as turning on a light [Jat *et al.*, 2019].

Thanu and Uthra [2015] have developed a software system that provides the ability to interact with home appliances using the Internet of Things concepts. This allows users to have flexible remote control mechanisms through a secure web connection to the Internet. Users are able to control appliances while away from home or monitor their status.

Folea *et al.* [2012] developed a smart home architecture in which they used a universal remote control that will allow the user to control the alarm, the lighting and the temperature in all rooms of the house. In addition, the system will be able to monitor dust in the air if windows and doors are closed.

In the paper [Schiel and Zitzelsberger, 2018] are presented 4 instruments for formant's detections: PRAAT tool by Boersma and Weenink [2017], SNACK (version 2.2) of the Snack Sound Toolkit of KTH Stockholm by Sjölander [2017], ASSP (version 2.8) contained in the Advanced Speech Signal Processor library by Scheffer [2017] and DEEP based on deep learning techniques and trained on corpus VTR-TIMIT [Dissen, 2016; Keshet, 2017]. The first three methods are based on LPC analysis and are language and corpus independent. DEEP has been trained on the part of VTR-TIMIT and is therefore language and corpus dependent.

In our paper we extract the feature vectors only from the vowel areas, and in order to make the segmentation of the consonant / vowel regions we used several detectors for fundamental frequency. In our future research, we intend to identify the interest areas with a VAD algorithm, which will help to the emotion recognition.

2 *The Fuzzy System for Emphasis of the Intonational Contours*

In this paper, the performance of emotion recognition with Deep-Learning neural networks was analysed, with the input set being spectral images emphasized in the areas of the F0 fundamental frequency and F1-F4 formants. The aim was to compare the accuracy of emotion recognition in Romanian, using the SROL emotion corpus, when emphasized MFSC images are used instead of standard MFSC images.

In order to identify the interest areas, we first looked for the location of the areas where F0 and F1-F4 formants are found, which are considered essential information in speech signals in terms of emotional load. Experienced persons in analysing and processing voice signals are able by looking at the spectrogram, based on pitch fluctuations (F0) and formants areas to visually identify the emotions expressed in the recording.

The Deep-Learning Neural Network (DL-CNN) with Convolutional Layers extracts its own set of features by applying in the input layers a gradual reduction of information through max-pooling layers, respectively applying various convolution masks that filter the signal. The applied convolution masks initially have random coefficients and depending on the characteristics of the applied LPF (Low-Pass Filter) or HPF (High-Pass Filter) filter, low or high frequency information is extracted. After a significant reduction of the input data achieved by successively passing through several input layers, the recognition is made by a finally full-connected MLP network, which associates different weights to the automated extracted feature vectors at the input layers. In general, DL-CNN can be considered as a "black-box" where we can vary the architecture of the network (number of layers, number of neurons per layer, neuron activation functions, etc.), but for the user, how was identified the valuable features for classification remains unknown.

In the signal spectrum obtained after applying FFT (Fast Fourier Transform) on the whole input speech signal, time domain information has been lost, so that for a dominant spectral component corresponding to a certain frequency it is no longer possible to specify in which time point it appeared in the initial signal. In the spectrogram, the spectrum is computed by on small analysis windows/frames of about 25 ms, with a window step of 10 ms, preserving the time domain information on the Ox axis and the frequency domain information on the Oy axis. In the MFSC images, the dominant spectral components such the fundamental frequency F0 and of the formants F1-F4 can be identified by the local maximum values and the moment of time when important variations of their values occurred is known. For a better detection of the F0, F1,..., F4 values are used spectral "envelopes" obtained by smoothing the cepstrum with different degrees (varying the size of the lifting window L) and a concatenation of the segments from these envelopes, where the useful information is found.

We developed a fuzzy-based algorithm to transform standard MFSC images in the idea that emphasizing the F0 and F1-F4 formant traces could increase emotion accuracy rate. The fuzzy system needs both minimum and maximum local values from the smoothed spectrum. We will note by mx_0 and mn_{01} the frequencies where the local minimum and maximum of the fundamental frequency F0 are found, respectively by νF_0 the spectral amplitude of this frequency. Similarly we will have for each formant F1,...,F4 extracted the values $\{mx_1, mn_{12}, \nu F_1\}, \dots, \{mx_4, mn_{34}, \nu F_4\}$. Assuming that we have multiple local minimum candidates between two consecutive values F0,F1,...,F4, then only the lowest minimum value will be kept and the rest will be neglected.

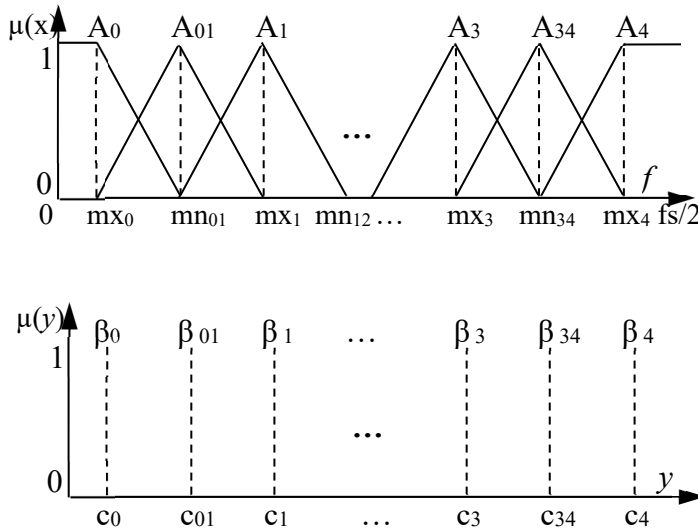


Figure 1 Input and output membership functions of the fuzzy system for emphasis of the intonational contours

The fuzzy system is of Sugeno type with a number of 9 triangular membership functions at the input. The singleton values at the output represent the degree of emphasis that is applied to each dominant formant value. The number of singletons may be identical to the number of input fuzzy membership functions or may be less if for several formants values identical degree of emphasis is applied. At the output we can have a minimum acceptable number of only two singletons, if for example in the local maxima of the formants F0, F1,...,F4 the same degree of emphasis is applied and in the local minima among them, the same factor would also be applied.

Table 1 Fuzzy rules of the fuzzy set of the emphasis intonational contour

X	$\tilde{A}_0$	$\tilde{A}_{01}$	$\tilde{A}_1$	...	$\tilde{A}_3$	$\tilde{A}_{34}$	$\tilde{A}_4$
Y	β_0	β_{01}	β_1	...	β_3	β_{34}	β_4

A fuzzy system was tested in which the rules had two premises (two input variables), but the results obtained were similar. This is the reason why we have chosen the fuzzy system with a single input variable.

IF x is A_i THEN y is β_i .

For computing the degree of membership for fuzzy sets A_{01} and A_1 the following calculation formulae is applied:

$$\begin{aligned} & \text{Equation 1} \\ \mu_{\tilde{A}_{01}}(f) &= \begin{cases} \frac{f - mx_0}{mn_{01} - mx_0} & , f \in [mx_0, mn_{01}] \\ 1 - \frac{f - mn_{01}}{mx_1 - mn_{01}} & , f \in [mn_{01}, mx_1] \\ 0 & , \text{otherwise} \end{cases} \\ \mu_{\tilde{A}_1}(f) &= \begin{cases} \frac{f - mn_{01}}{mx_1 - mn_{01}} & , f \in [mn_{01}, mx_1] \\ 1 - \frac{f - mx_1}{mn_{12} - mx_1} & , f \in [mx_1, mn_{12}] \\ 0 & , \text{otherwise} \end{cases} \end{aligned}$$

The membership degree values for the other input triangular fuzzy functions are similarly computed. Note that the values which define these triangular fuzzy functions are not fixed. Since for each 25.6ms analysis window/frame of the speech signal, a smoothed spectral envelope is determined, we will have differences between the formant values $F_0, F_1, \dots, F_4$ between two consecutive frames. The resulting spectrum for each analysis window represents one of the columns of the MFSC image and will have other local maximum and minimum values at the transition from one frame to the other. The triangular functions from the input of the fuzzy system will adjust for each individual frame.

The input functions have been specially chosen in such a way that we will have for any input value x we can have only a maximum number of two membership degrees and their sum is 1. Therefore except for the peaks of triangles and plateau areas $[0, mx_0]$ from the first input function, respectively $[mx_4, fs/2]$ of the last

fuzzy set, where only one fuzzy rule is activated, in all other cases only two fuzzy rules will be activated.

As an optimization to increase computational speed only one degree of membership μ_{Ai} can be computed applying Equation 1 and the second is obtained from the first:

$$\mu_{Ai,i+1} = 1 - \mu_{Ai}$$

For any input frequency the sum of membership degrees is 1. ($\sum \mu_{Ai}=1$)

IF x is A_i THEN y is β_i

IF x is $A_{i,i+1}$ THEN y is $\beta_{i,i+1}$

In each fuzzy rule, the values of the output singletons are truncated, according to the degree of the membership of the input, and because the fuzzy system has only one input variable (the rules have only one premise), there is no need to apply intersection operations between the premises such as in the situation when the fuzzy system had multiple inputs. The calculation of the output value of the fuzzy system, through the defuzzification operation, can also be optimized taking into account the previous observation:

$$y = \frac{\sum_t \mu_t \cdot \beta_t}{\sum_t \mu_t}$$

$$y = \frac{\mu_{\tilde{A}_i} \cdot \beta_i + \mu_{\tilde{A}_{i,i+1}} \cdot \beta_{i,i+1}}{\mu_{\tilde{A}_i} + \mu_{\tilde{A}_{i,i+1}}} = \frac{\mu_{\tilde{A}_i} \cdot \beta_i + (1 - \mu_{\tilde{A}_i}) \cdot \beta_{i,i+1}}{1}$$

The output value of the fuzzy system represents the emphasis degree applied by the multiplication operation to each frequency in the speech signal spectrum. If $y=1$, then the spectral magnitude does not change. If for example $y=0.8$, then the resulting value decreases to 80%, and if the value is greater than one e.g. $y=1.5$ then the amplitude increases by 50%. Because these changes are applied to the MFSC spectrogram pixels, it is required that the resulting values are in the range $[0, 255]$ with the maximum value of 255 corresponding to the white colour.

3 Algorithm to Emphasize the Spectral Images

A general algorithm of the proposed method is presented below:

- ```
for $k = 0 : 10\text{ms}$ (window step): length of signal
- computes the spectrum for the current analysis
 frame/window
- generates spectral envelopes with different degrees
 of smoothing using the signal cepstrum
- data fusion / fuzzy concatenation to get the final
 smoothed spectrum
```

## A NEW METHOD OF EMPHASIZING THE FORMANTS OF THE VOCAL SPECTRUM

```

- identification of local maximum and minimum values
 corresponding to dominant frequencies (on the basis
 of which input fuzzy sets are constructed)
for i = 1 to N (FFT - number_of_points)
 - apply Sugeno fuzzy system for current frequency
x = f(i)
 - after defuzzification the output of the fuzzy
 system $y = c$ represents the correction factor
 / degree of emphasis: $MFSC^*(i, k) = MFSC(i, k) \cdot$
 c
endfor
endfor

```

In general, in the speech signal spectrum in areas of interest represented by the vowels and semivowels, low frequencies have the highest amplitudes and high frequencies have significantly lower amplitudes. For this reason, in our simulations, we used an emphasis factor  $c=1.5$  (150%) for F0,  $c=0.7$  (70%) for the minimum values between the formants,  $c=1$  (100%) for F1 and F2 and  $c=2$  (200%) for F3 and F4.

These values specify the output singletons of the fuzzy system, but can be adapted according to the average of all formants  $vF_0, vF_1, \dots, vF_4$  and if, for example, it is desired to bring the dominant values in the spectrum to the same level, the emphasis factors are calculated according to the maximum mean value:  $c_i = vF_i / \max_k(vF_k)$ .

To test the effectiveness of MFSC images with emphasized intonational contours in automatic emotion recognition, a Deep-Learning neural network was used. The architecture of the network is shown in Fig. 2 and contains in the first input layers autoencoders that extract the important information for the classification and on the last layer a classical multi-layer full-connected MLP (Multi-Layer Perceptron) network that performs classification using the features automatically extracted at the input.

Note that the running times for such a network are about 2 hours, having ( $N_1=400, N_2=400, N_3=200$  neurons for autoencoding), and the running times decrease when the number of neurons is smaller. We also analysed how the recognition rate varies for DL-CNN architectures with fewer input layers and autoencoders with fewer neurons (when training times are lower), arriving at a training time that is about 15 minutes for an architecture  $N_1=100, N_2=100$ .

The MFSC images were built based on the files of the emotional database SRoL (Sound of Romanian Language) which contains 546 records with the four basic emotions (happiness-174, sadness-122, fury-163 and neutral tone-87).

The resolution of the images used is 41216 pixeli ( $\text{no\_features} \cdot \text{no\_frames} = 256 \cdot 161$ ), being about 3 times higher than the resolution of the standard MFSC images previously used where we had 13120 pixels. It should be noted, that some

records are shorter and a padding with zeros is performed so that all images in the dataset are the same size. The increase of image resolution in height was done to allow a better visualization of the intonational contour traces, but also required an increase of the number of layers in the DL-CNN architecture.

The training and testing sets contain 75% and respectively 25% of the 546 MFSC images, the distribution being random but keeping this percents for each emotion for all simulations. In order to report the results as accurately as possible, 10 simulations were performed (each with different training and test sets) and the results obtained for each simulation were averaged.

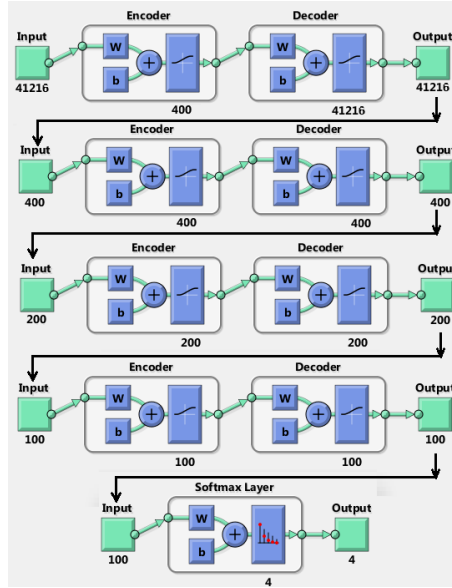


Figure 2 The Architecture of the DL-CNN neural network used

The DL-CNN architecture includes until 2 to 4 autoencoders, each autoencoder automatically extracts the relevant information from the input feature vectors and includes internal layers of convolution, max-pooling. The final layers are from a fully-connected neural network which classifies the inputs. The input MFSC images have the prosodic features (F1-F4 formants and F0 fundamental frequency) emphasized by the fuzzy algorithm.

## 4 Conclusions

DL-CNN provided slightly better results for emotion recognition when MFSC images with the emphasized intonational traces of F0 and F1-F4 formants than when standard MFCC images were used. The accuracy of emotion recognition for

the SRoL database was about 84% for normal MFSC images and about 86% for accentuated MFSC images. We expected to a slightly higher improvement in the recognition rate, however these are the best results obtained using spectral images with DL-CNN classification and we concluded that the input set of the emphasized MFSC is better.

## References

- Alghifari, M., Gunawan, T., Nordin, M., Qadri, S., Kartiwi, M., Janin, Z. (2019). On the use of voice activity detection in speech emotion recognition. *Bulletin of Electrical Engineering and Informatics*, volume 8, no.4, ISSN: 2302-9285, pp 1324-1332.
- Boersma, P., Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.23, retrieved 2017-04-17 from <http://www.praat.org/>.
- Dissen, A. (2016). Formant Estimation and Tracking using Deep Learning, In *Proceedings of the INTERSPEECH*, pp 958 – 962.
- Folea S., Bordencea D., Hotea C., Valean H. (2012). Smart home automation system using wi-fi low power devices. In: *Proceedings of 2012 IEEE International Conference on Automation, Quality and Testing, Robotics*, IEEE Press, Cluj-Napoca, Romania.
- Freeman, D.K., Cosier, G., Southcott, C.B., Boyd, I. (1989). The voice activity detector for the PAN-European digital cellular mobile telephone service. In: *International Conference on Acoustics, Speech and Signal Processing*, volume 1, pp 369–372.
- Gokul K., Lalitha S. (2018). Personality Identification Using Auditory Nerve Modelling of Human Speech. In *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1731-1737.
- Jat, D., Limbo, A., Singh, C. (2019). Voice Activity Detection-Based Home Automation System for People with Special Needs, Editor(s): Nilanjan Dey, *Intelligent Speech Signal Processing*, Academic Press, 2019, pp 101-111, ISBN 9780128181300, <https://doi.org/10.1016/B978-0-12-818130-0.00006-4>.
- Karray, L., Martin, A. (2003). Towards improving speech detection robustness for speech recognition in adverse environment. *Speech Communication*, 40 (3), pp 261–276.
- Keshet, J. (2017). DeepFormant, retrieved 2017-04-30 from <https://github.com/MLSpeech>.
- Ong, W. Q., Tan, A.W.C. (2016). Robust voice activity detection using gammatone filtering and entropy. In: *International Conference on Robotics, Automation and Sciences (ICORAS)*, pp 1-5.
- Ramirez, J., Segura, J., Benitez, C., Torre, A., Rubio, A. (2004). Efficient voice activity detection algorithms using long-term speech information. *Speech Communication*, vol. 42, pp 271–287, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.9761&rep=rep1&type=pdf>.
- Sangwan, A., Chiranth, M.C., Jamadagni, H.S., Sah, R., Prasad, R.V., Gaurav, V. (2002). VAD techniques for realtime speech transmission on the Internet. In *IEEE International Conference on High-Speed Networks and Multimedia Comm.*, pp 46–50.
- Scheffer, M. (2017). Advanced Speech Signal Processor (libassp), retrieved 2017-04-17 from <http://www.sourceforge.net/projects/libassp>.

- Schiel, F., Zitzelsberger, Th. (2018). Evaluation of Automatic Formant Trackers, In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), pp.2843-2848, <https://aclanthology.org/L18-1449.pdf>.
- Sehgal A., Kehtarnavaz N. (2018). A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. In IEEE Access, volume 6, pp 9017-9026.
- Sjölander, K. (2017). Snack-Sound-Toolkit, retrieved 2017-04-17 from <http://www.speech.kth.se/snack>.
- Sohn, J., Sung, W. (1998). A voice activity detector employing soft decision based noise spectrum adaptation. In International Conference on Acoustics, Speech, and Signal Processing, vol. 1, pp 365– 368.
- Song, J. (2018). Research on Digital Hearing Aid Speech Enhancement Algorithm. In 37th Chinese Control Conference (CCC), pp. 4316-4320.
- Thanu, S.S., Uthra A.R. (2015). A smart home energy management system in intelligent building. Int. J. Eng. Dev. Res. 3 (2), pp 400–403.

# Romanian Text Anonymization Experiments from the CURLICAT Project

VASILE PĂIȘ, ELENA IRIMIA, RADU ION, DAN TUFIȘ,  
MARIA MITROFAN, VERGINICA BARBU MITITELU,  
ANDREI-MARIUS AVRAM, and ERIC CUREA

*Research Institute for Artificial Intelligence "Mihai Drăgănescu",  
Romanian Academy*

{vasile, elena, radu, tufis, maria, vergi, andrei.avram,  
eric}@racai.ro

## Abstract

Considering both Romanian and international regulations with regard to protecting personal information, anonymization has become an important processing step before publicly releasing any text corpora. In this paper, we describe different experiments performed in the context of the CURLICAT project towards anonymization of large corpora. The resulting algorithms were implemented as a tool, which is publicly available and also integrated in the RELATE platform for online usage. Furthermore, we provide a brief overview of the future CURLICAT corpus, which will be anonymized using the algorithms and tools presented in the paper.

## 1 Introduction

The General Data Protection Regulation (GDPR)<sup>1,2</sup> clearly states that "the protection of natural persons in relation to the processing of personal data is a fundamental right". The regulation contains provisions and requirements for handling personal data of individuals, whatever their nationality or place of residence, and applies to any enterprise established in the European Economic Area (EEA). One of the aims of this regulation is to give individuals control over their personal data. For this purpose, different concepts need to be considered, such as: right to be forgotten, privacy by design [Spiekermann, 2012], transparency (full information is provided to individuals) and accountability (demonstrable responsibility for using personal data) [Goddard, 2017].

All automatic processes must take into consideration protecting personal information by employing techniques such as anonymization. Data anonymization is the process of removing personally identifiable information from collected data. Article 4, paragraph 5, of the GDPR, defines "pseudonymisation" as the

---

<sup>1</sup> Council Regulation 2016/679, 2016 O.J. (L119) (EU) 1

<sup>2</sup> <https://eur-lex.europa.eu/eli/reg/2016/679/oj>

processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information. Francopoulo and Schaub [2020] consider that fully anonymized data that meet the legal bar set by the European data protection law is no longer personal data and is therefore not subject to the obligations of GDPR. Furthermore, the authors distinguish between "local" pseudonymization, where a pseudonym is used within a single text to replace certain personal information, and "global" pseudonymization, where the same identifier is used in multiple texts to replace the same personal information. In the case of "global" pseudonymization, some re-identification may still be possible by combining different information from different contexts. Depending on the available information, such as social network information, mobility traces, epidemiological and health-care data, de-anonymization attacks are possible as discussed in [Ji *et al.*, 2017]. From this perspective, "local" pseudonymization seems more resilient against de-anonymization attacks since there is less context available from which to infer personal information.

Following the success of the project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL)<sup>3</sup>, where a large legislative corpus [Váradi *et al.*, 2020] was constructed, comprising monolingual corpora in 7 EU languages, including Romanian [Tufis *et al.*, 2020], we went on to create a new large resource relevant to Connecting Europe Facility (CEF) Automated Translation (AT) eTranslation platform, covering domains relevant to all Digital Service Infrastructures (DSIs). Thus, in the context of the project "Curated Multilingual Resources for CEF.AT" (CURLICAT)<sup>4</sup>, the consortium partners aim to deliver language resources, in particular monolingual corpora, in the EU/CEF languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian. All targeted seven EU's official languages are under-resourced languages and consequently, languages without sufficient technological support. The monolingual corpora will represent the following specific domains: finance, health, scientific research, cultural heritage, education, economics, politics. Even though the data will be cleared from the point of view of Intellectual Property Rights (IPR), there remains the problem of protecting potentially sensitive personal data.

In the context of the CURLICAT project, we developed several anonymization algorithms for Romanian language corpora. The aim is to provide a high-level of pseudonymization, in compliance with existing regulations, while still allowing the corpus to be useful for language research and training automatic natural language processing applications. In this work, we present the different

---

<sup>3</sup> <https://marcell-project.eu/>

<sup>4</sup> <https://curlicat-project.eu/>

experiments that were performed as well as give preliminary information with regard to the Romanian CURLICAT corpus.

The rest of the paper is structured as follows: Section 2 provides a short description of the Romanian CURLICAT corpus, Section 3 presents related work, Section 4 defines the anonymization requirements from the point of view of the present work, in Section 5 are presented the proposed algorithms, followed by implementation details in Section 6. Finally, conclusions are given in Section 7.

## 2 *Short Description of CURLICAT Romanian Corpus*

The Romanian component of the CURLICAT corpus was constructed based on the The Reference Corpus of the Contemporary Romanian Language<sup>5</sup> (CoRoLa [Mititelu *et al.*, 2018]). Documents in CoRoLa can be classified according to different metadata attributes, from which relevant in our selection were DocumentTextStyle (with the values Journalistic, Science, Administrative, Imaginative, Memoirs, Law, Other), DocumentTextDomain (with the values Arts and Culture, Nature, Society, Science, Other) and DocumentTextSubdomain, with 65 possible values, associated with only one possible domain. Documents with the DocumentTextStyle=Imaginative were not included in the selection because belletrist texts cannot be assigned a domain and, therefore, they are not of interest for the CURLICAT project. Values of the DocumentTextDomain and DocumentTextSubdomain attributes were mapped to the CURLICAT targeted domains and this mapping was used to extract documents corresponding to the following: Culture, Economy, Education, Health, Nature, Politics and Science.

CoRoLa is an IPR-cleared corpus in the sense that RACAI, as one of its developers, has rights to download, store, process and make available for search all the texts in the corpus. In the context of the CURLICAT project, we obtained from the data providers the right to distribute selected texts to the Directorate-General for Translation (DGT) via the ELRC-Share platform. The targeted size of the selected corpus is at least 2 million sentences, i.e., approximately 37 million words. Since we extracted data for 7 CURLICAT domains, to obtain a balanced corpus we estimated 300,000 sentences for each domain. Statistics of the selected IPR-cleared data are presented below. As can be seen in Table 1, the target of 300,000 sentences was covered for all domains and even exceeded at this moment, but we intend to work on balancing the data and therefore the delivered size will be drastically reduced in the Science domain (which is too broad), and adjusted in the Culture and Health domain.

After selection, the data went through a further process of clearing, to automatically reject sentences written in foreign languages or the ones containing a lot of abbreviations, numbers or other symbols. The cleaned corpus was processed in the RELATE platform, using the integrated UDPipe [Straka and

---

<sup>5</sup> <https://corola.racai.ro>

Straková, 2017]. Processing involved segmentation, tokenization, lemmatization, part of speech tagging, dependency parsing. The generated files are in CoNLL-U Plus format<sup>6</sup>. Statistics were computed using the RELATE platform and certain relevant indicators are given in Table 2.

*Table 1 Statistics for the number of sentences in the CURLICAT-Ro corpus*

| CURLICAT domain | No. of sentences |
|-----------------|------------------|
| culture         | 577,307          |
| education       | 320,484          |
| economy         | 311,721          |
| health          | 417,681          |
| nature          | 338,953          |
| politics        | 379,188          |
| science         | 2,113,454        |
| TOTAL           | 4,458,788        |

*Table 2 Statistics on the processed files*

| Indicator                | Value      |
|--------------------------|------------|
| Number of Documents      | 29,010     |
| Number of Tokens         | 121,882,49 |
| Hapax Legomena           | 685,90     |
| Dis Legomena             | 149,99     |
| Number of Unique Lemmata | 1,016,984  |
| Number of Unique Tokens  | 1,237,119  |
| NOUN                     | 23,739,27  |
| PROPN                    | 4,923,21   |
| PRON                     | 5,870,00   |
| VERB                     | 9,746,96   |
| NUM                      | 2,633,19   |
| ADJ                      | 7,090,11   |
| ADV                      | 5,153,49   |

### 3 Related Work

As the legal framework covering privacy regulations evolved, both in Europe and abroad, there has been equal consideration from the research community working in the field of language technology. This can be seen from the series of Legal Issues workshops, organized in recent years [Choukri *et al.*, 2020] with support from the European Language Resources Association (ELRA) and associated with the International Conference on Language Resources and Evaluation (LREC). Other workshops dedicated to more specific topics have been organized as well, such as the Workshop on NLP and Pseudonymisation [Ahrenberg and Megyesi, 2019].

Already in 1996, Sweeney [1996] employs templates and specialized knowledge of what constitutes a name, address, phone number and so forth to detect and replace what is considered personal information. Dates were replaced

<sup>6</sup> <https://universaldependencies.org/ext-format.html>

with a similar date nearby, while personal names were replaced with a fictitious unique name that sounded reasonable. A similar approach is described in [Douglass *et al.*, 2004]. Dates are shifted by a random number while trying to keep the days of the week, while personal names are replaced with names from a publicly available list, randomizing first and last names. Similarly, locations were replaced with randomly selected small town names and hospital and clinical units were given fictitious names.

Chevrier *et al.* [2019] conducts a review of 60 full-text articles from the MEDLINE<sup>7</sup> bibliographic database published between 2007 and 2017 discussing de-identification or anonymization. MEDLINE is the National Library of Medicine's (NLM) premier bibliographic database and the primary component of PubMed<sup>8</sup>. The authors of the review conclude that there is a growing interest for privacy-enhancing methods in the life sciences community that crosses scientific boundaries, involving primarily computer science, biomedical informatics and medicine.

Dalianis [2019] employs a rule-based algorithm for anonymizing electronic patient records in the Swedish language. Name lists, composed of Swedish person names, locations, town squares, parks, postal codes, area codes and prefixes for phone numbers, are used to identify potentially sensitive information. Once identified, information is replaced with similar surrogates. For example, in the case of personal names the replacement is chosen from a list of the 100 most frequent Swedish names.

Di Cerbo and Trabelsi [2018] propose a supervised machine learning system to identify personal information (such as name, address, age). The authors acknowledge the shortcomings of using regular expressions due to the different forms in which addresses and other types of personal information can be written. However, since regular expressions are proven to be very good at recognizing certain types of data following strict patterns (such as emails and IP addresses), the authors propose to combine multiple machine learning algorithms for named entity recognition and regular expression-based matching.

Meystre *et al.* [2018] acknowledges the time consuming nature of manual anonymization of patient data and proposes three types of automatic systems: a) rules-based, using pattern matching augmented by a false positives filter, b) using Conditional Random Fields (CRF) [Lafferty *et al.*, 2001] and c) an ensemble combining four different machine learning algorithms. Their findings suggest that the ensemble system outperforms the CRF-based one which in turn outperforms the rules-based system.

Francopoulo and Schaub [2020], while focusing on a Customer Relationship Management (CRM) system, makes use of a previously developed, commercially

---

<sup>7</sup> <https://www.nlm.nih.gov/medline/index.html>

<sup>8</sup> <https://pubmed.ncbi.nlm.nih.gov/>

available, named entity recognition (NER) system, based on the output of a syntactic parser [Francopoulo, 2008] and a cascade of pattern matching rules. The authors acknowledge that the quality of the overall process is highly dependent on that of the NER system. For proper names, language-based local clues are combined with a large list of names extracted from WikiData<sup>9</sup>.

#### 4 *Anonymization Requirements in the CURLICAT Project*

Before implementing an anonymization solution, the type of sensitive information to be protected must be identified. Given the domains they belong to, there is a very low probability that the documents in the CURLICAT corpus contain IP addresses, emails, phone numbers, bank accounts or any personal identifiers (such as personal numeric codes). Therefore, for the anonymization experiments presented in this work we did not consider such personal information.

In the case of the CURLICAT corpus, most anonymization requirements appear in relation to news and other blog posts. In this case, the entities to consider are represented by person names, locations and organizations. If person names are obvious candidates for anonymization, the requirements for locations and organizations may need some clarification. The GDPR only applies to personal data, belonging to an individual. However, in certain cases, by maintaining location or organization names it is possible to infer the person's name. For example, considering expressions such as "primarul orașului X" ("the Mayor of X city") combined with the knowledge that this is a news article and the approximate period for the article it can be deduced who the person is. Similarly, given the expression "proprietarul hotelului Y" ("the owner of hotel Y"), assuming it is a small hotel and there is no other hotel with the same name in Romania then the person's name can be inferred.

Expressions such as "hotelul Y se afla în orașul X" ("the hotel Y is situated in the city X"), taken alone, do not require anonymization, since there is no personal information involved. However, given the context, even this apparently harmless sentence can be used to deduce personal information. For example, if the sentence is followed by another sentence starting with "Proprietarul său ..." ("Its owner ..."), then again, the observations made previously apply and the name of the owner can be inferred. Therefore, we adopt a strict approach and anonymize all instances of persons, organizations and locations.

Identified entities must be replaced with a pseudonym or identifier, thus removing the personal data from the text. We agree with Francopoulo and Schaub [2020] and consider that "local" pseudonymization provides better protection compared to "global" pseudonymization. For this purpose, we develop algorithms that work at document level and not at corpus level. Therefore, each document will be anonymized individually providing identifiers valid only in the document

---

<sup>9</sup> [https://www.wikidata.org/wiki/Wikidata:Main\\_Page](https://www.wikidata.org/wiki/Wikidata:Main_Page)

context. Furthermore, we take into account the existence of large documents. For this purpose, the algorithm can be applied at sub-document level, thus reducing the context in which a certain identifier represents the same entity. From a technical perspective, by "local" pseudonymization it becomes possible to apply the algorithms in parallel across multiple documents or document blocks without the need of synchronization between processes. Each anonymization process is in charge of a single text fragment and is not aware of other anonymization operations within the same corpus.

As described in Section 3, some authors decided to replace identified personal information with "similar" information extracted from entity lists (for example, Dalianis [2019] employs a list of the most frequent 100 Swedish names). For the purposes of the CURLICAT project, we decided to employ identifiers that would ensure anonymization while allowing future corpus users to replace the anonymized information with something that may be more suitable for their respective works. It is envisageable that for training certain information extraction systems it may be needed to replace identified entities with different forms (single word, multiple words, frequent names, less frequent names, etc.). In addition, we were concerned with allowing future users to construct words appropriate for a certain context. Thus, considering the proper name "Maria" with the additional form "Mariei", we can imagine contexts such as "Casa Mariei este scumpă" ("Mary's house is expensive") and "Casa Maria este scumpă" ("The Mary House is expensive"). In the first case the form "Mariei" expresses the ownership of the house, while in the second case the form "Maria" indicates that it is probably a guest house, which from the point of view of named entities is an organization with the name "Casa Maria".

The CURLICAT corpus will be delivered in two formats: raw anonymized text and token-based annotations. However, the token-based annotated format must contain anonymized text as well. This means the anonymization algorithm has to be either applied twice (once on the raw text and once on the annotated documents) or the identifiers used must be resilient to the annotation procedure. In the first case, two algorithms must be actually developed (one working on the raw text and one working on the tokenized documents). However, in the second case, the identifiers that were used as an alternative to the personal data must behave like words in order to allow the annotation algorithm to properly mark them with the corresponding part-of-speech tags and any additional information. Furthermore, any segmentation and tokenization algorithms must recognize the identifiers as a single block and make sure that it is mapped to a single token.

## 5 Algorithms

Given the requirements defined in Section 4, we decided to implement "local" pseudonymization on raw text. To account for different forms of writing a name, the identifiers used will contain the suffix as well (when a suffix is being used).

Furthermore, the identifier should remain unchanged during tokenization and other tagging operations. For this purpose, we experimented with several identifier variants and arrived at the following form: `_#TYPE#ID_suffix`, where *TYPE* is the named entity type, *ID* is a number associated with the entity and *suffix* is the suffix. For example, "Maria" becomes `_#PER#1_`, while "Mariei" becomes `_#PER#1_ei`. By keeping the suffix, we allow future users of the corpus to replace the identifiers with similar word forms using list of names. In this case `_#PER#1_` could be replaced with "Elena" and `_#PER#1_ei` could be replaced with "Elenei", thus allowing other algorithms to function on text similar to the original one, while ensuring proper anonymization.

1. For each n-gram (n=10..1),  
     if found in the NE list  
     generate an ID for the appropriate class
2. If not found:
  - 2.1. try constructing an entity by adding/removing suffixes  
     (for example: Mariei -> Maria, Clujului ->Cluj)  
     and search in NE list
  - 2.1.2. If found, generate an ID for the appropriate class
3. If not previously found:
  - 3.1. If the word starts with uppercase:
    - 3.1.1. Check a 1-gram model trained on CoRoLa for frequencies with  
 uppercase and lowercase starting letters
    - 3.1.2. If `freq(lowercase)>freq(uppercase)` then Ignore the word
    - 3.1.3. Else generate an ID for the UNK class

*Listing 1 Algorithm based on name lists and frequencies*

We conducted several experiments with texts containing the proposed identifiers, tokenized and annotated using UDPipe [Straka *et al.*, 2016]. Each time the tokenizer recognized the identifier as a single token. Furthermore, when the suffix was present, the lemmatizer attempted to derive a correct form. Considering the previous example `_#PER#1_ei` a lemma was proposed having the form `_#PER#1_a`, while for the case of not having a suffix the lemma was identical to the word form. Considering the part-of-speech tags, the identifier was marked as proper noun (PROPN).

The first experiment performed made use of a rule-based method for matching names against name lists. For this purpose, we made use of name lists previously

created<sup>10</sup> in the ReTeRom project. The lists contain 15,944 person names, 5,336 locations and 6,441 organizations. Since certain less-used names may not be on the list, we further introduced an unknown (UNK) entity type. In this case, we identify words beginning with a capital letter and check for the frequency of usage in lowercase form using a frequency model computed on the CoRoLa corpus [Tufiş *et al.*, 2019]. The algorithm is presented in Listing 1.

In addition to the name lists algorithm, we used a named entity recognition system [Păiş *et al.*, 2021] implemented using the LegalNERo gold corpus [Păiş *et al.*, 2021]. This is a manually annotated corpus extracted from the Romanian sub-corpus [Tufiş *et al.*, 2020] of the larger MARCELL corpus [Váradi *et al.*, 2020]. The LegalNERo corpus comes with two annotation layers. The first one contains named entities for persons, organization, locations, time and legal references. The second layer contains only persons, organizations, locations and time annotations. This accounts for sub-entities (mostly organizations and time expressions) present in the legal reference entities. For the purposes of this work, we used the second annotation layer, which contain entities relevant for the anonymization process. For each of the identified entities we then generate an identifier according to the rules expressed above. The algorithm is presented in Listing 2.

1. Execute NER system
2. For each identified entity of relevant type
  - 2.1. Generate an ID for the appropriate class,  
taking into account the suffix (if present)

*Listing 2 Algorithm based on NER system*

Finally, the two algorithms from Listing 1 and Listing 2 were combined. The overall F1 score associated with the NER system is 86.84%. For many applications this score is good enough. However, for the purposes of anonymization we considered that a higher recall is beneficial in order to make sure no personal data remains in the final corpus. For this reason, the NER system was first executed and then the name lists algorithm was applied. Finally, the rules from Listing 2 were also applied to mark unknown entities that potentially still contain personal information.

## 6 Implementation

The algorithms were implemented as a Java application, exposing the functionality in the form of REST APIs. This means an application needing to anonymize a text can make a HTTP call with the document to the server and will

---

<sup>10</sup> <https://racai.ro/p/reterom/rezultate.html>

receive a JSON answer containing the anonymized text as well as the correspondences between identifiers and removed text spans. It is up to the final application to keep or discard the correspondences, in compliance with the usage scenarios and regulations.

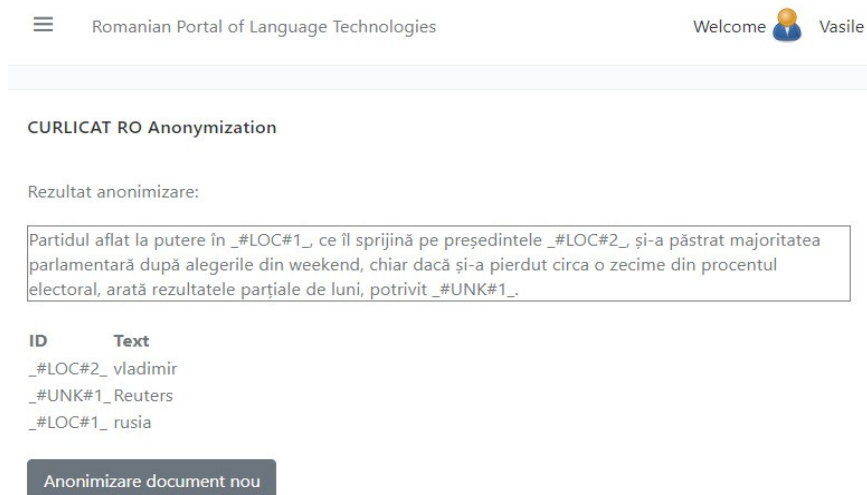


Figure 1 Example integration in the RELATE platform

From an implementation point of view, whenever possible, we used efficient data structures, such as Hash Maps, ensuring quick response time (a hash map structure has a complexity  $O(1)$  for lookups). This provided an average processing speed of 8,026 tokens/second for the algorithm presented in Listing 1 (without the NER system) and 206 tokens/second when using the combination of algorithms from Listing 1 and 2 (executing also the NER system). For computing the processing speed, we used an Intel Xeon Silver 4210 CPU running at 2.2Ghz and we averaged the speed obtained over 10 tests on the same documents.

The NER system was constructed in Python using recurrent neural networks, based on BiLSTM cells, with a final CRF layer. For features we considered word representations, character embeddings, gazetteer resources and known affixes. The word embeddings are initialized from a pre-trained model and are fine-tuned during training. The character embeddings are computed during training and the embedding layer is followed by a BiLSTM layer that helps with the representation generation. The resulting NER system is also implemented in the form of a REST API server. It receives the text document and returns the identified annotations in the form of text spans with associated entity type. The anonymization application acts as a client for the NER system, invoking the NER REST endpoint with the received document and then making use of the provided output. This follows best

practices related to web service reusability [Henderson and Yang, 2004] and further allows for a dynamic architecture by specifying the NER endpoint as a parameter at the anonymization application start-up time.

The resulting application was integrated in the RELATE<sup>11</sup> platform [Păiș *et al.*, 2019; Păiș *et al.*, 2020]. This integration allows an end-user to perform text anonymization using a web-based graphical user interface (GUI). The GUI is presented in Figure 1. The user can enter the desired document and then is presented with the anonymized text together with the list of identifiers and their corresponding text spans.

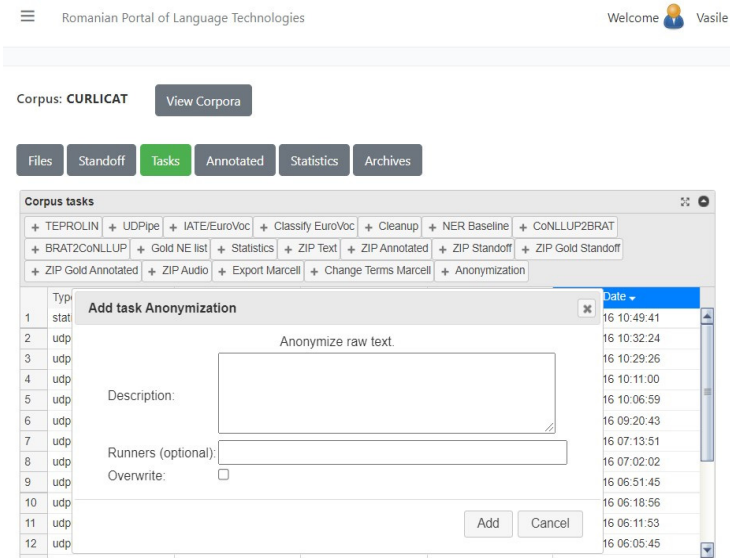


Figure 2 Anonymization task in the RELATE platform

A similar integration was realized in the corpus management area of the platform. This allows large corpora to be anonymized directly from the platform by adding an "anonymization" task. In order to speed-up processing of large volumes of text, multiple processes can be started and the platform automatically balances the calls between the processes. This allows parallelization of the workload, by anonymizing simultaneously multiple files. The task creation interface is presented in Figure 2. When creating a new task, the user has the ability to specify the processes used to parallelize the execution (by default all the available processes are used). Furthermore, the platform automatically makes sure

<sup>11</sup> <https://relate.racai.ro>

only new files, not yet anonymized, are being processed by the new task. However, the user can re-process all the files by selecting the "overwrite" option.

## 7 Conclusion and Future Work

In this paper, we presented several anonymization experiments for Romanian language and the system that will be used to anonymize the CURLICAT corpus. This final system combines the output from a NER system with a rule-based method. Since neither of the two components is perfect, we focused on reducing the false negatives (entities that are missed and remain present in text). Therefore, we consider that our approach increases the anonymization quality of the system, while still allowing future natural language processing tasks to be applied successfully, due to inclusion of suffixes in the pseudonymization identifiers. The resulting system is available for direct usage within the RELATE platform and the source code is made available on our GitHub<sup>12</sup>.

Since the NER component is integrated as a web service, we plan to test other systems as well. Furthermore, we plan to prune the rules to provide better results. The primary goal is to reduce the false positives, thus keeping more data, that does not require anonymization, in the final documents.

## Acknowledgements

This work was conducted in the context of the "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project, CEF-TC-2019-1 – Automated Translation grant agreement number INEA/CEF/ICT/A2019/1926831.

## References

- Ahrenberg, L., Megyesi, B. editors. (2019). *Proceedings of the Workshop on NLP and Pseudonymisation*, Turku, Finland, September. Linköping Electronic Press.
- Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., Lovis, Ch. (2019). Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. *Journal of Medical Internet Research*, 21(5): e13484, May.
- Choukri, K., Linden, K., Rigault, M., Siegert, I. editors. (2020). *Proceedings of the LREC2020 Workshop on Legal and Ethical Issues*. European Language Resources Association, Marseille, France, May.
- Dalianis, H. (2019). Pseudonymisation of Swedish electronic patient records using a rule-based approach. In *Proceedings of the Workshop on NLP and Pseudonymisation*, pp 16–23, Turku, Finland, September. Linköping Electronic Press.
- Di Cerbo, F., Trabelsi, S. (2018). Towards personal data identification and anonymization using machine learning techniques. In András Benczür, Bernhard Thalheim, Tomáš Horváth, Silvia Chiusano, Tania Cerquitelli, Csaba Sidló, and Peter Z. Revesz, editors,

---

<sup>12</sup> [https://github.com/racai-ai/ROAnonymization\\_CURLICAT](https://github.com/racai-ai/ROAnonymization_CURLICAT)

- New Trends in Databases and Information Systems*, pp 118–126, Cham. Springer International Publishing.
- Douglass, M., Clifford, G.D., Reisner, A., Moody, G.B., RG, M. (2004). Computer-assisted de-identification of free text in the mimic ii database. In *Computers in Cardiology, 2004*, pp 341–344.
- Francopoulo, G., Schaub, L.P. (2020). Anonymization for the GDPR in the Context of Citizen and Customer Relationship Management and NLP. In *workshop on Legal and Ethical Issues (Legal2020)*, pp 9–14, Marseille, France, May. LREC2020, ELRA.
- Francopoulo, G. (2008). TagParser: well on the way to ISO-TC37 conformance. In *International Conference on Global Interoperability for Language Resources*.
- Goddard, M. (2017). The EU General Data Protection Regulation (GDPR): European Regulation that has a Global Impact. *International Journal of Market Research*, 59(6):703–705.
- Henderson, P., Yang, J. (2004). Reusable web services. In *Jan Bosch and Charles Krueger*, editors, *Software Reuse: Methods, Techniques, and Tools*, pp 185–194, Berlin, Heidelberg, Springer Berlin Heidelberg.
- Ji, S., Mittal, P., Beyah, R. (2017). Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: A survey. *IEEE Communications Surveys Tutorials*, 19(2):1305–1326.
- Lafferty, J., McCallum, J., Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pp 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Meystre, S., Heider, P., Kim, Y., Trice, A., Underwood, G. (2018). Clinical text automatic de-identification to support large scale data reuse and sharing: Pilot results. In *Proceedings of the AMIA 2018 Annual Symposium*.
- Barbu Mititelu, V., Tufiş, D., Irimia, E. (2018). The reference corpus of the contemporary Romanian language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Păiş, V., Tufiş, D., Ion, R. (2019). Integration of Romanian NLP tools into the Relate platform. In *The 14th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 181–192.
- Păiş, V., Tufiş, D., Ion, R. (2020). A processing platform relating data and tools for Romanian language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 81–88, Marseille, France, May. European Language Resources Association.
- Păiş, V., Mitrofan, M., Gasan, C.L., Coneschi, V., Ianov, A. (2021). Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.
- Spiekermann, S. (2012). The challenges of privacy by design. *Commun. ACM*, 55(7):38–40, July.
- Straka, M., Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99.

- Straka, M., Hajic, 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'16)*, pages 4290–4297, Portorož, Slovenia, May. European Language Resources Association (ELRA).
- Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. *Proceedings of the American Medical Informatics Association AMIA Annual Fall Symposium*, pp 333–7.
- Tușiș, D., Barbu Mititelu, V., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little strokes fell great oaks creating Corola, the reference corpus of contemporary Romanian. *Revue Roumaine de Linguistique*, 64(3):227 – 240.
- Tușiș, D., Mitrofan, M., Păiș, Ion, R., Coman, A. (2020). Collection and annotation of the Romanian legal corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp 2773–2777, Marseille, France, May. European Language Resources Association.
- Váradi, T., Koeva, S., Yamalov, M., Tadic, M., Sass, B., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tușiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J. (2020). The MARCELL legislative corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France, May. European Language Resources Association.

## **CHAPTER 5**

### **LANGUAGE VARIETIES AND SOCIAL MEDIA**



# Statistics of Some Regional Varieties of Romanian Greetings

ANCA-DIANA BIBIRI<sup>1</sup>, MIHAELA COLHON<sup>2</sup>,  
and MIHAELA MOCANU<sup>1</sup>

<sup>1</sup>*Department of Social Science and Humanities,  
Institute of Interdisciplinary Research,  
"Alexandru Ioan Cuza" University of Iași*

<sup>2</sup>*Faculty of Computer Science, University of Craiova*

{anca.bibiri, mcolhon, mihaelamocanuiasi}@gmail.com

## Abstract

The scope of the present study is to provide an image about lexical diversity of Romanian greetings and farewells. Seen as an expression of politeness and as a means of sociolinguistic integration, with an important role in organizing the act of communication, greetings fulfil some specific functions: they mark the beginning or the ending of a communication act; they initiate verbal interaction with the interlocutor; they indicate the speaker's communication intentions; they express the speakers' social status; they signal the relationships between the participants in the communication act. In 2015-2017 a corpus consisting of greeting formulae recorded in the dialectal sociolinguistics surveys was created in the project *The Contrastive Analysis of Romanian and Spanish Intonation* (SoRoEs). At the methodological level, the statistical analysis of the corpus data was performed using several software tools for identifying the sequence patterns that characterize greetings in our language and for manipulating and visualizing the corpus contents. Taken into considerations the diatopic dimension, our results display important differences across regions regarding the most used formulae: *bună!* in Muntenia, *salut!* in Moldova, and *servus!* in Transilvania. Taken into consideration three sociolinguistic variables: age, gender and level of education, our analysis reveals that more differences were recorded at the subjects' level of education and age, while gender does not display significant changes, but reflects different degrees of emotional load.

## 1 Introduction

A corpus, defined as a significant collection of authentic language raw text no matter how large is it will obviously be of finite size. As a direct consequence, it cannot contain the occurrences of all possible linguistic phenomenon, because corpus linguistics use finite portions of language use with the assumption they will reveal some aspect of a given language [Desagulier, 2017: 3]. Corpora generally cannot record all instances of a language or a subset thereof. In particular, the

totality of general language cannot be known and as a consequence cannot be captured in its entirety [Breyer, 2011: 22]. The incompleteness of linguistic data in any corpus impose to the scientific works to build general models from incomplete data. In corpus linguistics, the term is used differently – it refers to a collection of language expressions that must ensure the following properties:

- the instances contained in the corpus must be authentic, which means they must include only real-life instances;
- the collection is representative for the language;
- the collection is large [Stefanowitsch, 2020: 22].

A legitim question arises: *How large a linguistic corpus must be?*” Taking into account the fact that the size of a corpus is relative, a possible answer could be that it must be large enough to contain a sample of instances of the phenomenon under analysis. It is well known that any text, spoken or written, will lose not only its communicative context (the discourse that generated it) but also some of its linguistic and paralinguistic properties when it becomes part of a corpus. This is more obvious in case of transcribed spoken data (as is the case of this corpus) where the very act of transcription means that aspects like tone of voice, intonation, subtle aspects of pronunciation, facial expressions, gestures, etc. are replaced by flat textual sequences. In addition, the texts of such collections are usually annotated in order to enhance their potential for linguistic analysis [Stefanowitsch, 2020: 39].

We have ensured the corpus under analysis is authentic as it is made of authentic language expressions as they were all extracted from communicative situations. These situations were created from conversational acts, more precisely interviews between real people: one who conducted the discussion being in charge with the corpus realization and the respondents from 10 cultural cities across Romania. It also covers the most popular greeting and farewell formulae used in all Romanian regions, excepting the religious formulae that made the topic of other research/study.

The way the corpus was designed and realized by collecting “real life” language examples, two main aspects of any well-done corpus was ensured: data reliability, in terms of how accurate the data was collected, and data validity, in terms of how representative the selected subjects are and how faithful the recorded samples are to the variability that characterizes the respective language [Desagulier, 2017: 3]. Typically, for a corpus to be representative (or “balanced”) the corpus creators must include different manifestation of the language by keeping also into account the right proportions, which reflect the corpus utterances in the community in question. This means that the corpus contents must reflect both quantitatively and qualitatively the language forms or language varieties exactly as they appear in the real-life communities. Obviously, this is an ideal target impossible to be

achieved in reality. To support this idea, we have to point out that even the communities are not homogeneous, so defining a corpus by taking equal proportions of the language varieties in each community may not yield to a realistic language balanced representation.

Given the issues discussed above it seems impossible to create corpora that meet both the representativeness and balancedness criteria. We can only approximate representativeness and balancedness of languages in linguistic corpora, as they represent an ideal than an attainable goal [Leech, 2007].

There is another aspect about the corpus data: as it only contains the recorded utterances given in their textual form, it needs further annotations to provide information about the semantics or pragmatics features of the contained linguistic expressions. For cognitive linguistics with corpus data, it needs some interpretation annotations in order to provide meanings to the data. In the absence of such information, we can only analyze the formal patterns of the linguistic expressions for greetings and farewells.

Software packages for corpus analysis allow to query the corpus for particular linguistic expressions or abstract data related to the information represented in the corpus contents. According to Stefanowitsch [2020] most of these software packages have the following functions, specifically:

- to generate *concordances* which means word forms that occur in the proximity of a certain language constructions by offering an overview of the typical usage of a target set of word forms;
- to identify *collocates* of a given expression, which means word forms that occur in certain position relative to a target expression/word and in this way the contexts of a linguistic expression can be summarized;
- to produce *frequency lists*, meaning lists of the corpus contents in terms of character strings together with their frequency of occurrence by offering information about the distribution of the word forms [Stefanowitsch, 2020: 22].

Thus, linguistic phenomena in a corpus are examined taken into consideration concordances, collocations, and frequency lists, providing useful information about the typical usage as greetings and farewells in our language. In this sense, Stubbs noticed that corpus investigations often lead to discovering that "absolutely fixed patterns are extremely rare, and a frequent conclusion is that a given pattern is 'typical' or 'canonical' but that it has variants" [Stubbs, 2009: 118], conclusion confirmed also by our analysis.

The article is organized as follows: in the Related Works section, the state-of-the-art updates our understanding of statistics in corpus linguistics. Methodology section discusses the approach of our corpus in the context of the state-of-the-art. The following section displays/discusses the results of our statistical processing

analyses. We conclude our paper with the interrelationship between some sociolinguistic variables via correlation.

## 2 *Related Work*

The present study is a predictable sequel of Mocanu and Bibiri [2019] in which the corpus contents are statistically measured using frequency and distribution data. The scope of the present study is to provide an image about lexical diversity of Romanian greetings and farewells. Word or token frequencies as a dimension of lexicographic characteristics are important in many studies because of their correlations with many experimental tasks. For this main reason, frequency is often used as a control variable in statistical analyses of corpora.

We tested several statistical methods based on the sociolinguistic variables depicted to be characteristic for the corpus contents (see [Mocanu and Bibiri, 2019; Mocanu, 2018]). They are considered as potentially relevant parameters as they are equally distributed in the population being demographic variables (e.g., age, gender, and level of education).

As an essential communicative skill, greetings are language- and culture-specific [Duranti, 1997]; in this sense, studies about regional variation in greetings in Germany are presented by Schneider and Baron [2008], also Wierzbicka [1985] discusses greetings in Polish and Australian English; also, Pinto [2008] contrasted the use of greetings in peninsular Spanish with US English underling the differences concerning politeness strategies; while Schüpbach [2014] compared Swiss German with Germany German. Sociolinguistics variables as age and gender are taken into account by Moradi [2018] who compared how these two sociolinguistics variables affect the greeting behavior in Persian and English, and draws the conclusion that the interlocutors' age affects greeting comportment in similar ways in both languages, but that gender influences the greeting choices more in Persian.

Nilsson *et alii*, within the framework of variational pragmatics, investigate with interactional and statistical analyses to what extent the different variables affect variation in greeting forms: "While nation indeed is an important factor, the study also illustrates how social variables such as age, gender and participant roles as well as situational variables such as medium, region and venue impact the greeting choices participants make" [Nilsson *et al.*, 2020: 1].

As it is well known, a corpus does not bring new information about language, but the software offers us a new perspectives on the familiar [Hunston, 2002: 3]. Thus, the software employed to analyse corpora count frequency and search collocations: corpus linguistics typically works with numbers which reflect the frequencies of words and phrases in corpora [McEnery and Hardie, 2011]. In such a way, statistical analysis "help us discover and elucidate patterns and tendencies in the data that might otherwise remain hidden" [Brezina, 2018: 5].

### 3 Methodology

At the methodological level, the statistical analysis of the corpus data was performed using several software tools for identifying the sequence patterns that characterize greetings in our language and for manipulating and visualizing the corpus contents. The implemented statistical analysis is visually shown with line graphs and Boxplots. Interrelationship between some sociolinguistic variables via correlation are also investigated. For this scope, Pearson's correlation was determined by taking into account the scale variables while Spearman's correlation was conducted on ordinal variables.

It is well known that corpus linguistics is a distributional discipline been among the most quantitatively oriented linguistic disciplines. For this reason, distributional frequency data is usually investigated in such collections. We consider here the approach that corpus linguists deal more with statistical tools than with language processing ones. After all, statistics is the scientific discipline that deals with quantitative distributions. Despite that, not all corpus-based studies utilize the available statistical methods to their fullest extent.

In this study we search for distributional data across the three sociolinguistic variables annotated in the corpus: age, gender and level of education. More precisely we generate frequencies of occurrence of the linguistic elements of the corpus, the Romanian greetings expressions, from several different perspectives:

- which are the most frequent greetings in each of the three main Romanian historical regions, such as in *Transylvania*, *Muntenia* and *Moldova*. This information is provided in various different forms of frequency graphs as it is shown in Figures 5-10;
- how evenly the greetings constructions are distributed across the participants in the questionnaire by taking into account their gender (Figures 11 and 12), educational level (Figures 13-15) and age (Figures 16-18)
- how evenly the formal Romanian greetings<sup>1</sup> are distributed across the participants in the questionnaire by considering all the three sociolinguistic variables (Figure 19).

The implemented statistical analysis is visually shown with bar and pie graphs. Interrelationship between some sociolinguistic variables via correlation are also investigated (as in Figure 19).

#### 3.1 Greetings Corpus

Our corpus of greetings and farewells was obtained through a combination of traditional methods of guided conversation with a Discourse Completion task.

---

<sup>1</sup> These constructions are widely spread in the Romanian population being the most used formal greetings: *Bună ziua!*, *Bună dimineața!* and *Bună seara!*.

They were recorded based on a questionnaire containing 13 open-answer discussion instances of communication. This is aimed at identifying the greeting/farewell formulae used by subjects in various situational, formal and informal situations. In this respect we provided a short description of the situational context for each respondent, without suggesting any possible answer or expression choice, as illustrated below (with the possible answers expected from the subjects according to the contextual or pragmatic conditions):

- You run into a colleague on the street: *Bună!!* ‘Hi!’ (other formulae: *mneăța!!* ‘morning’, *salut!!* ‘hello’, *bonjour!*, *bonsoir!*, *servus!*, *ciao!*), Mihaela/Mihai! ‘You look very nice! Where are you going?’ – ‘What a beautiful dress! Where did you get it from?’ (to a man:) ‘You look very smart! Where are you going? You must go on a date.’
- In situation when you are together with a friend whom you introduce to another person: ‘Let’s have coffee’; ‘Let’s have a glass of wine’; ‘Let’s have a toast!’
- You run into a friend whom you haven’t seen in a long time: ‘Is it you, Mihaela/Mihai? I barely recognized you; it’s been so long! How are you? What have you been doing? Where do you live now?’
- Visiting your relatives: ‘Good morning, mother/father!’, ‘Good afternoon!’; (other greetings: *Săru-mâna!!* ‘I kiss your hand!’ *Mă bucur să te văd!!* ‘Glad to see you again!’, *Bine v-am găsit sănătoși!!* ‘Glad to find you all well!’);
- You meet one of your former teachers: *Bună ziua, domnule profesor!!* ‘Good afternoon, Mr. Professor!’
- In a public institution you ask for some documents from a clerk: *Bună ziua!!* ‘Good afternoon!’ (if the interlocutor is a man), and *Sărut mâna!!* ‘I kiss your hand!’ (if the interlocutor is a lady);
- Entering a compartment on the train: *Bună ziua! Este vreun loc liber?!* ‘Good afternoon! Any seats available? Do you mind if I sit there?’
- A GP appointment: *Bună dimineața, domnule/doamna doctor!!* ‘Good morning, Mr./Mrs. doctor!’
- You are late for work and your boss asks to see you. You enter his/her office: *Bună dimineața, domnule director! Mă scuzați că am întârziat!!* ‘Good morning, Mr. Director! I’m sorry for being late!’

Thus, we obtained a Mini\_RoGreet corpus that comprises, besides proper/well-known greetings and farewells, also collocations and greeting-questions that are

seen as more powerful and requests for information (for more details see [Bibiri *et al.*, 2016]).

### **3.2 *Subjects and Points of Survey***

All the participants were native speakers of the Romanian language who were living at the time of recording in the following 10 cities selected across the country in order to illustrate the standard literary Romanian spoken language: Baia-Mare, Braşov, Bucureşti, Cluj, Constanţa, Craiova, Iaşi, Sibiu, Suceava and Timişoara. The respondents were native speakers of the Romanian language and covered all three social variables taken into account: age, gender and level of education. The respondents must have met the classic criteria required in a sociolinguistic dialectal survey: must have originated in the points of investigation, communication availability, spontaneity, and good diction.

#### **3.2.1 *Processing of Recorded Data/ Recordings and Acoustic Analysis***

The surveys were recorded with a Marantz Professional recorder (PMD661MKII), in digital format (having extension .wav – Waveform Audio File Format). Recordings were made in a quiet environment without background noise (as much as context/surroundings permits so), and the sound wave has a frequency of 48 KHz. The acoustic analysis was performed with software tools: GoldWave and Praat. The selected sequences were aligned as follows: delineation and labeling according to the statements used in the questionnaire, coded in order to respect the GDPR rights.

#### **3.2.2 *Corpus Analysis***

The implemented statistical analysis is visually shown with line graphs and Boxplots. Interrelationship between some sociolinguistic variables via correlation are also investigated. Most of the data in this paper was analysed using quantitative methods that looked at the frequency and distribution of different greetings and farewell words.

The corpus analysis was implemented in Python language. The corpus was managed for analysis using *pandas*, a Python library for working with tabular data. The corpus contents were added to DataFrame objects and then the implemented statistics were graphically illustrated using *matplotlib* library. We also used *wordcloud* library in Python in order to visualize some greetings formulas together with their frequencies in corpus.

The results of this analysis upon the Romanian greeting and farewells corpus (Mini\_RoGreet) is discussed below.

#### 4 *Discussion and Results*

In our analysis we have recorded 1085 greetings formulas (140 unique) divided approximatively equally between men and women respondents (as it can be seen in Figure 1). More precisely, 507 greetings were spoken by men while 577 greetings were recorded from women respondents.

Also, we obtained a corpus that has balanced proportions regarding the participants' age (see Figure 3), with a less dimension for the young respondents: we have recorded 263 greetings from people having between 25 and 35 years old, 406 greetings from the next category (with age between 36 and 45 years) and 416 greetings from people having between 46 and 55 years old.

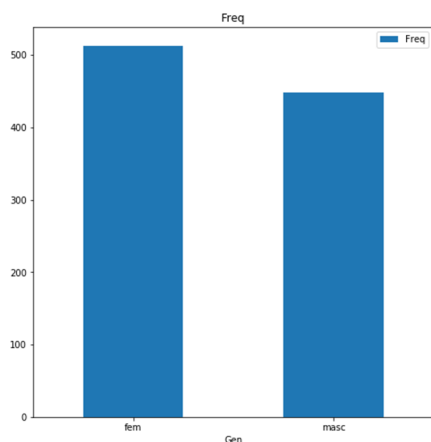


Figure 1. Population grouped according to their gender

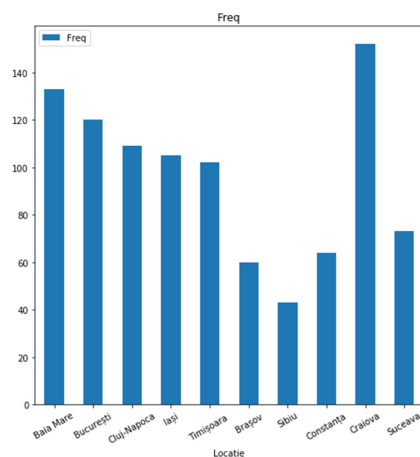


Figure 2. Population grouped according to their location

Regarding the respondents' location, as far as that goes, due to the fact that in Sibiu we recorded only 4 subjects, here we get the smallest number of greetings and farewell formulae. While Craiova represents the top of it, according to a great number of interviewed subjects, and therefore the largest number of greeting and farewell formulae (see Figure 2).

These graphs underline the sociolinguistic variables taken into account when recording the corpus. As one can notice, there are almost equal proportions concerning gender, and in a small degree, age, while the level of education is unbalanced (due to the fact that very few interviewees fall into the first category of studies).

As it can be seen from the following three Figures (5, 6, and 7), the greetings formulae used in all Romanian regions are: the formal ones, as the triadic system,

## STATISTICS OF SOME REGIONAL VARIETIES OF ROMANIAN GREETINGS

corresponding to the three periods of the day: *Bună dimineața!*/ *Bună ziua!*/ *Bună seara!*, *Sărut-mâna!*, *La revedere!*, and the informal ones: *Ciao!*, *Pa!*/ *Pa, pa!*, *Bună!*, *Salut!*/ *Salutare!*, *Am plecat!*, *Bine ați venit!*/ *Bine ai venit!*, *Bine v-am găsit!*, *O zi bună!*, *Te pup!*, *Hai salut!*, *Mneața!*/ *Neața!*, *Să trăiți!*, *Hai noroc!*, *Ne vedem!*/ *Ne vedem mâine!*, *Sănătate!*, *Hei!*, *Am venit. Ce (mai) faci?/ Ce (mai) faceți?* (this kind of question-greetings were discussed in [Bibiri *et al.*, 2016]).

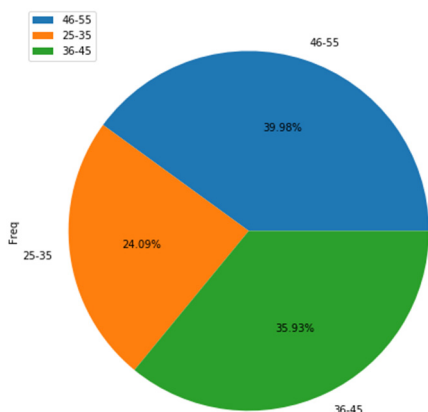


Figure 3. Population grouped according to their age

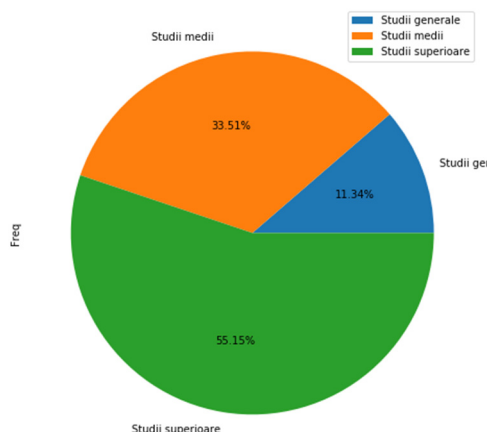


Figure 4. Population grouped according to their studies

According to Figure 5, the most used informal greeting formula in Muntenia region that opens the conversation is *Bună!*/‘Hi!’, and for leaving, very often is used *Pa!*/‘By!’ while the preference for *Salut!*, *Ciao!* that cancel the opposition between greetings and farewells, being used both at the beginning and at the ending of a communicational act:

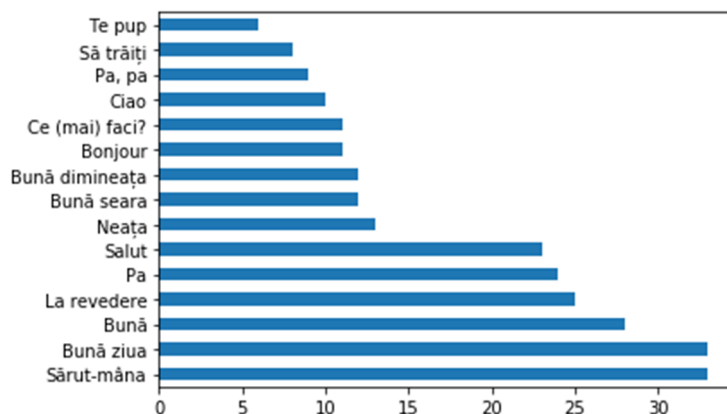


Figure 5. Most frequent greetings in Muntenia country

In Figure 6, we caught the most frequent formulae that subjects greet each other in Transilvania region: in addition to the popular Romanian greetings such as *Buna ziua!*, *Salut!*, *La revedere!*, at the informal level, by far, *servus!* is on top, followed by *salut!* and *ciao!* (with its variants *ceau!*, *ciau!* – originated from Italian). It is this tendency that the common Romanian greetings corresponding to the triadic phases of the day to be replaced by *servus!*

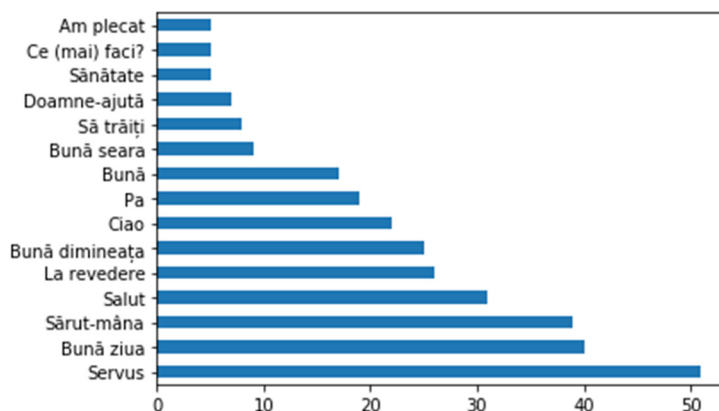


Figure 6. Most frequent greetings in Transilvania country

In what concerns the Moldavian region, representing the informal register, *salut!* stands out from all the other greeting formulae, followed by *bună!*, and, by now, long-standing and used greeting-question *Ce mai faci?*. While *pa!* (reiterated *pa-pa!*) is preferred by subjects as leave-taking formula, followed by *Am plecat!* and *Ne auzim!*. As a greeting used more in formal context, *sărut-mâna!* (with the

version *sărut-mâinile!* – used exclusively by men to show profound respect for a woman) occupies the first place showing a high degree of deference.

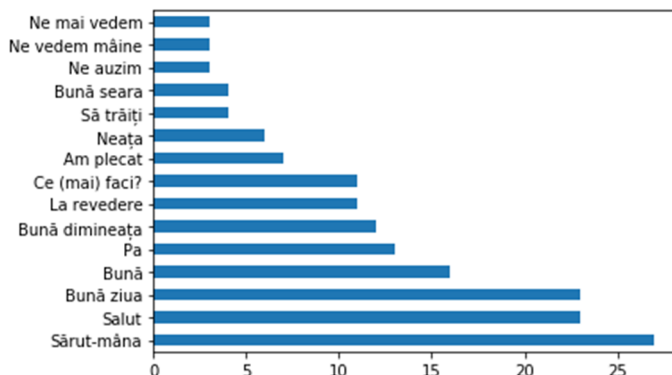


Figure 7. Most frequent greetings in Moldova country

Expressing the highest courtesy, *Sărut-mâna!* has an interesting approach: addressing older relatives (be they mother, grandmother, aunt, godmother or father, grandfather, godfather); when it comes about the hierarchic relationships, addressed to women who are professionally superior (employer-boss; student-teacher); in other formal context (as for example, patient-doctor/clerk/teacher relationship); when addressing to the officials of the church (although the gesture accompanying the greet tends to disappear even in rural area); nowadays, this formula is used with the meaning of *mulțumesc!*, expressing gratitude for a favour, and also after meals.



Figure 8. Transylvania greetings-cloud

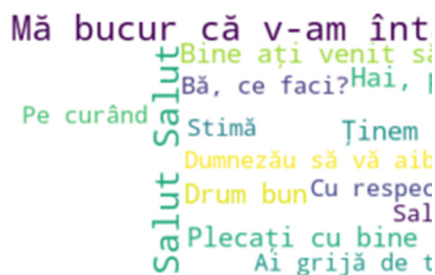


Figure 9. Moldova greetings cloud

At the diatopic level one may notice differences registered by the greeting formulae, at the standard language level, most of them specific to the Transylvania



## STATISTICS OF SOME REGIONAL VARIETIES OF ROMANIAN GREETINGS

that are used by a category, but this does not imply the fact that the other category does not use them (see Figures 11 and 12).

More frequent usage of one form by one of the genders might indicate either that the form is ‘gender preferential’, and to some extent whether it is a standard or vernacular form, as women tend to use more standard language and men more vernacular forms [Holmes, 2008:163].

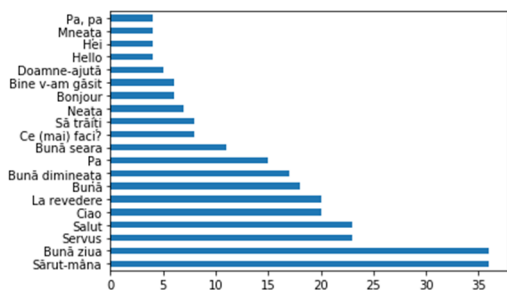


Figure 13. Greetings of general education people

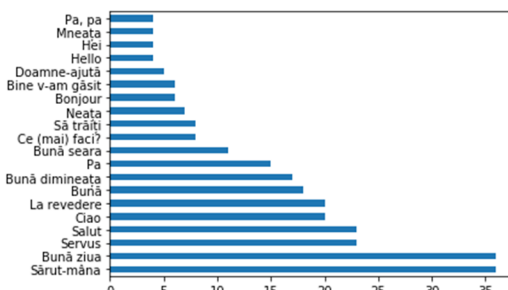


Figure 14. Greetings of secondary education people

Greetings/farewells exclusively used by men are: Să trăiți!, Să trăiești!, Noroc! (‘May you live long!’, ‘Cheers!’), while women are preference for: Te pup! / ‘Kiss you!’, ‘Lots of love!’, Pa, pa! / ‘Ta-ta!’, Hai, pa! / ‘Bye, now!’

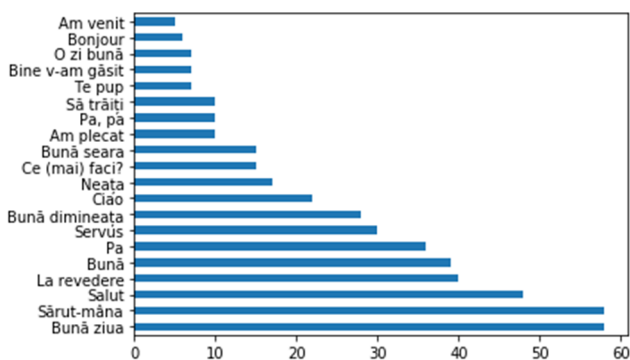


Figure 15. Greetings of higher education people

The level of education has a significant impact on the choice of greeting and farewell formulae: the most used are formal, official, neutral formulae (among people of unequal social status): *Bună dimineața!*, *Bună ziua!*, *Bună Seara!*, *La*

*revedere!*, *Sărut mâna!*; while the most used informal, non-conventional formulae are: *Salut!*, *Servus!*, *Ciao!*, *Pa!*, *Hai, pa!*, *Hai, noroc!*, *Salutare!*, *Bună!*, *Hei!*.

It is worth mention that the formula *hei!* ‘hey!’ (Figure 16) used only by young people the age category of 25-35 years old is a more recent phenomenon, that expressed joy and closeness (although it is not a proper greeting), and it is used in informal contexts. However, it is also possible/predictable that young people using *hey!* as a greeting is language change in progress [Holmes 2008: 216]. Taking into account that *hey!* displays a limited frequency we conclude that it is used vernacular as lower classes use more vernacular forms [Coulmas, 2007: 28].

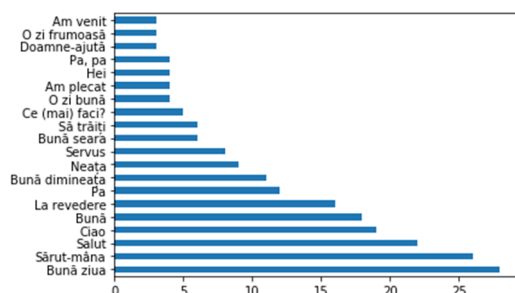


Figure 16. Greetings of 25-35 people

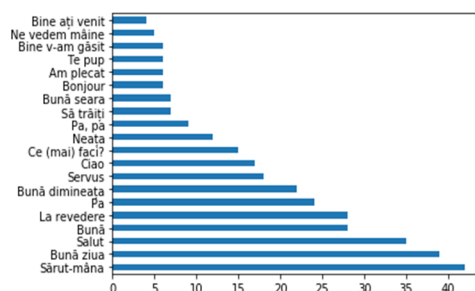


Figure 17. Greetings of 36-45 people

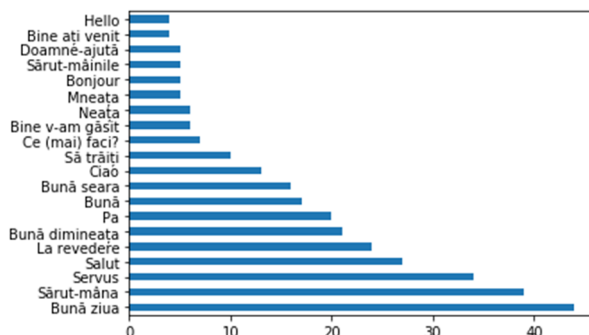


Figure 18. Greetings of 46-55 people

We generate an overall chart for all official greetings of Romanian language such as: *Bună ziua!*, *Bună dimineața!* and *Bună seara!* (see Figure 19). From the resulted graph we can conclude that from the “general education” people, these greetings formulas are used mostly by women of “46-55” age-category. For the

“secondary education” people, the age area of people that use these greetings include even the “35-46” age-category, but remains popular between women of “46-55” age-category. For “higher education” category, these formulas are common greetings regardless their age or gender.

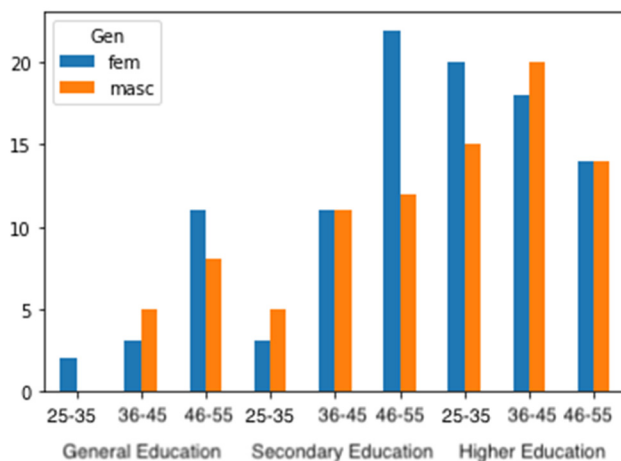


Figure 19. The attributes of the people that use official Romanian greetings

## 5 Conclusions

Language use is in constant flux [Baker, 2010], thus Romanian greetings and farewells present a dynamic paradigm resulting from the interconnection of older forms or combinations of them, and the borrowings from other language (facilitated by migration and the new media) as an effect of the globalization process.

Our statistical analysis concerning the regional distribution of Romanian greeting shows that there are canonical patterns for different regions of our country: Transylvania is individualized by the use of *servus!* with its variants: *serus!*, *serbus!* (and *servustok!* as plural form) that enjoyed the privileged status displaying the highest frequency, followed closely by *ciao!*. In Muntenia, *bună!* is far more frequent than other greeting words, followed by *salut!* Moldova region is represented by the French greeting word *salut!* used alternatively with *bună!*. These patterns are used both when people meet and when they part.

This study provides empirical evidence to existing theories on greetings and seeks to clarify within what contexts and different social groups, various greetings are used and seen as normative or become so. The choice of greeting to some extent can be a key to determine a person's age, gender and social class, or indicate respect or solidarity. The data in this study found statistically significant differences in the choices of greeting people make based on context, class, gender and age.

The outcomes of this research concerning the level of education underline the fact that greetings used in formal situations are the official ones and more used in the same degree: *bună ziua!*, *sărut-mâna!*, employed by unequal social status of the interlocutors. When leaving, the most common word is *la revedere!*, perceived as being polite, even a distant farewell. While in informal situations, the repertoire of formulas used by interlocutors of the same social status is rich and very diverse: *servus!*, *salut!*, *bună!*, *ciao!*, *noroc!*, even the very used greeting-question *Ce mai faci?*.

Regarding the age, formal greetings such as: *bună ziua!*, *bună dimineața!*, *bună seara!*, *sărut-mâna!* are employed both at the beginning and at the end of verbal interaction, and they were recorded for all age categories. While young people represent the most original and inventive segment of population, and are prone to use greetings and farewells borrowed from other languages, especially English, in order to express a modern attitude and to stand out from the other community members and belongingness to a certain social group.

According to speakers' gender, there are formulae used only by men: *Să trăiești!*, *Să trăiești!*, *Noroc!*, and greetings and farewells formulae, generally attributed to women: *Te pup!*, *Pa-pa!*, *Hai pa!*, who shows a higher degree of affectivity, whereas in the case of men, affectivity is conveyed by gesture, postures and attitudes in order to express various meanings in transmitting the message.

This study opens further directions in investigated greetings and farewell formulae from different perspectives and also those used especially in social-media platforms where communication is shorter and accompanied or better replaced by more visual icons. All these will give a complete picture about the politeness ritual in contemporary Romanian language.

## References

- Bibiri, A., Mocanu, M., Turculeț, A. (2016). The Intonation of Romanian Greetings: a Sociolinguistics Approach. *International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, World Academy of Science, Engineering and Technology*, 10 (4): 653-661, <https://doi.org/10.5281/zenodo.1123715>
- Alexandra Breyer, Y. (2011). *Corpora in Language Teaching and Learning: Potential, Evaluation, Challenges*, Peter Lang.
- Brezina, V. (2018). *Statistics in Corpus Linguistics. A Practical Guide*. Cambridge University Press.
- Coulmas, F. (2007). *Sociolinguistics. The Study of Speakers' Choices*, Cambridge, Cambridge University Press.
- Desagulier, G. (2017). *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. Springer International Publishing AG.
- Duranti, A. (1997). Universal and culture-specific properties of greetings, *Journal of Linguistic Anthropology* 7(1): 63–97, <https://doi.org/10.1525/jlin.1997.7.1.63>

- Gries, S. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, 3: 1225-1241.
- Hunston, S. (2002). Corpora in Applied Linguistics. *Cambridge University Press*.
- Leech, G. (2007). New Resources, or Just Better Old Ones? The Holy Grail of Representativeness. In *Corpus Linguistics and the Web*, ed. Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer, 133–149. Amsterdam: Rodopi, [https://doi.org/10.1163/9789401203791\\_009](https://doi.org/10.1163/9789401203791_009).
- McEnery, A., Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge, New York: Cambridge University Press.
- Mocanu, M., Bibiri, A.D. (2019). Greetings and Farewells in Contemporary Romanian: a sociolinguistic approach. *Zeitschrift für romanische Philologie (ZrP)*, 135 (3): 866-886, <https://doi.org/10.1515/zep-2019-0046>.
- Mocanu, M. (2018). The Greeting in the Romanian Language. Gender Differences in the Use of Greetings. In *Philologica Jassyensia*, 14 (1): 83-95.
- Moradi, R. (2017). Sociolinguistic aspects of the speech act of greeting in Persian and English. *Bulletin de la Société. Royale des Sciences de Liège* 86, 297-303. <https://doi.org/10.25518/0037-9565.6724>.
- Nilsson, J., Norrthon, S., Lindström, J., Wide, C. (2018). Greetings as social action in Finland Swedish and Sweden Swedish service encounters – a pluricentric perspective. *Intercultural Pragmatics*, 15: 57-88. <https://doi.org/10.1515/ip-2017-0030>
- Pinto, D. (2008). Passing greetings and interactional style: a cross-cultural study of American English and Peninsular Spanish. *Multilingua*, 27: 371-388. <https://doi.org/10.1515/MULTI.2008.017>
- Schüpbach, D. (2014). German or Swiss? Address and other routinised formulas in German-speaking Switzerland. *Challenging the Monolingual Mindset*, 156, 63. <https://doi.org/10.21832/9781783092529-007>
- Silverstein, M. (2003). Indexical order and the dialectics of sociolinguistic life. *Language Communication* 23 (3-4): 193-229.
- Stubbs, M. (2009). The search for units of meaning: Sinclair on empirical semantics. Memorial article. John Sinclair (1933-2007). *Applied Linguistics*, 30 (1): 115-137. <https://doi.org/10.1093/applin/amn052>



# Offensive Language Identification in Social Media for Romanian Language

MIHAELA ONOFREI<sup>1,2</sup>, DIANA TRANDABĂȚ<sup>2</sup>,  
and BOGDAN-ANDREI DONU<sup>2</sup>

<sup>1</sup>*Institute for Computer Science, Iași branch of the Romanian Academy*

<sup>2</sup>*Faculty of Computer Science, “Alexandru Ioan Cuza” University of Iași,*

mihaela.plamada.onofrei@gmail.com,

dtrandabat@info.uaic.ro, donu\_bogdan\_andrei@yahoo.com

## Abstract

The number and size of social networks have increased significantly over the years. With 1.5 billion active users, Facebook is by far the most popular social network on the planet. Our aim in this paper is to show that people tend to use social media in order to express more negative points of view, rather than positive ones. Facebook and other online platforms strive to automatically identify comments with inappropriate language in order to keep their users safe from cyberbullying. To solve this problem, intelligent systems are needed that process this information quickly and automatically signal potential threats, to limit the number of messages sent to be verified by moderators. In this article, we analyzed a series of comments randomly collected from the online environment. The developed models were evaluated and the logistic regression model, based on TFIDF features, is considered to perform best on this task, with a correct classification rate of aggression cases of approximately 87.0%.

## 1 Introduction

Offensive language, also referred to as cyberhate, is generally defined as the use of “violent, aggressive or offensive language” [Castano-Pulgarín *et al.*, 2021], mostly directed towards an individual or groups, having as a trigger different matter of religion, gender or political beliefs. The International Convention on the Elimination of all forms of Racial Discrimination called for a ban on expression of racial superiority or inferiority [Gagliardone *et al.*, 2015]. The Internet’s speed, the widely use of social networks, their “free” language and the virtual environment added to the natural tendency of humans to be rather critical than pacifist determined UNESCO to make a report that provides a glimpse of the dynamics characterizing offensive language in social media and suggested a few measures that could prevent this phenomenon. [Gagliardone *et al.*, 2015] showed that 80% of European Union citizens dealt with offensive language and 40% of them felt threatened via Social Media Networks. Social deviance is a concept that can illustrate offensive language, it encompasses all kinds of behaviors, from the

least violent to those that are considered law-breaking [Henry, 2009], from defamation, to call for violence, to displaying discriminatory views [Bhavnani *et al.*, 2009]. Due to these issues, the automatic identification of offensive language is important in order to automatically remove the offensive comments and posts.

In this context, we started analysing social media behaviour for Romanian language. We therefore developed a manually annotated corpus of offensive language, and tested two classification methods in order to investigate the usage of cyberhate in Romanian.

The rest of the paper is organised as follows: Section 2 discusses different state of the art systems for offensive language identification, Section 3 presents the corpus and the overall architecture of the system, Section 4 shows the obtained results, before summing up in the final conclusion section.

## 2 *Similar Approaches*

Discussions about analysing comments and other articles in online environments in order to detect feelings conveyed in social media messages are extensive in the literature, especially in the last decade. Castano-Pulgarín *et al.* [2021] offer an extensive analysis over the online offensive speech, covering 67 studies which took place between 2015 and 2019. Among them, Ventirozos and collaborators [2018] claim that the typical methodology for detecting cyber harassment on social media has two phases. It extracts general characteristics of keywords (text-based), characteristics for feelings that are rare in other contexts (network-based), but are frequently expressed in bullying posts, along with characteristics that describe the author's profile for each message (user-based). Using these extracted features, the message can be classified as showing an aggressive language or not.

Prior to any type of data analysis, it is necessary to prepare and clean them, and in the case of comments apply tokenization, PoS labeling (part of speech), lemmatization [Van Hee *et al.*, 2018] or stemming [Rasel, Sultana, Akhter, & Meesad, 2018], after which one can move on to the actual analysis. There are several methods that can be used in analyzing comments, obtaining different degrees of correct classification, among which we mention Random Forest, MaxEnt, TF-IDF and Hidden Markov models.

Through TF-IDF (word frequency tracking) and Random Forest to analyse whether a comment is offensive or not, obtaining an accuracy of 93% [Rasel *et al.*, 2018] while a Hidden Markov classifier had an accuracy of only 58.63% in classifying whether a comment falls into the category of bullying or not [Ventirozos *et al.*, 2018]. Other researchers have used MaxEnt models for which they found that they had a greater ability to discriminate posts from Random Forest, demonstrating that MaxEnt can be successfully adopted in cyberbullying studies with unbalanced data sets [Dadvar and Niamir, 2016].

Because we have these different results regarding the methods of analysis, it is advisable to use several classification methods, which we usually compare by the percentage of correct classification.

### **3 *Corpus Description***

Most offensive language detection systems work for English, but we aimed at developing a system for Romanian. Thus, our first task was to create a corpus of social media texts and annotate it for offensive language. Using two web scraping tools, DataMiner and Comment Exporter, 2.156 comments were extracted from Facebook and from two news forums (<https://www.g4media.ro/> and <https://www.hotnews.ro/>). The total number of words in all comments was 10561. From these total words, we removed duplicates and kept only verbs, nouns and adjectives, obtaining a corpus of 4686 words for which it was possible to associate positive and negative scores using RoWordNet [Dumitrecu, 2015].

Among the collected comments, a number of 450 comments were manually annotated, 274 comments as offensive and 176 as non-offensive. During the annotation phase, we noticed that opinions are expressed in very varied ways by different users, however annotator intra- and inter-agreement was about 90%.

### **4 *System Description***

The architecture of our system follows the general approach, starting with corpus cleaning and pre-processing (tokenization, PoS labeling, lemmatization), performing sentiment analysis and finishing with classification. We considered two methods for classifying comments as offensive or not: the first method uses a composed positive/negative score of the comment, and the second one uses logistic regression.

Each step is shortly described and exemplified below.

#### **4.1 *Data Cleaning***

This step uses regular expressions to remove all URLs (e.g., [www.xyz.com](http://www.xyz.com)), hashtags (e.g., #topic), targets (@username), punctuation, symbols, numbers and stopwords. Additionally, acronyms are expanded using a dictionary of acronyms.

#### **4.2 *Tokenization and Lemmatization***

Each post or comment is stored in a string. To allow further processing, this string was divided into individual words and the base-form of each word was identified using the NLPtools available from the Faculty of Computer Science of Iasi.

| Input                                                                 | Output                                                             |
|-----------------------------------------------------------------------|--------------------------------------------------------------------|
| In ritmul asta vom fi tigru <span>l</span> Europei....la covid<br>😂😂😂 | ['ritmul', 'tigru <span>l</span> ', 'europei', 'covid', '']        |
| I-au furat pe Donnie. Care i-ati furat bai like-urile lui Trump ?     | ['I', 'furat', 'donnie', 'furat', 'bai', 'like', 'urile', 'trump'] |

Figure 1 Data cleaning example

| lema                     | pos       | pos_type    | clean            | length |
|--------------------------|-----------|-------------|------------------|--------|
| d                        | NOUN      | common      | d                | 1      |
| k                        | NOUN      | common      | k                | 1      |
| f                        | NOUN      | common      | f                | 1      |
| e                        | VERB      | predicative | ie               | 1      |
| ă                        | VERB      | predicative | ă                | 1      |
| lema                     | pos       | pos_type    | clean            | length |
| fbclidiwarfxkdisonfgw... | ADJECTIVE |             | fbclidiwarfxk... | 46     |

Figure 2 Lemmatization example

After lemmatization, we noticed that a number of words were either too short (1 letter) or too long (more than 40 letters). Thus, we removed them also, and finished having 4686 different lemmata.

### 4.3 Parts of Speech

Parts of speech, such as adjectives, adverbs and some groups of verbs and nouns are good indicators of subjectivity and feeling. Negation is also an important feature, since the presence of a negation usually changes the polarity of the expressed opinion. Using the same NLPTools, the corpus was annotated with part of speech information.

| Input      | Output                                                                                                                                                              |
|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| esti       | <W LEMMA="esti" MSD="Np"<br>POS="NOUN" Type="proper" id="1.17" offset="117">esti</W>                                                                                |
| mincinosul | <W Case="direct" Definiteness="yes" Gender="masculine"<br>LEMMA="mincinos" MSD="Afpmsry" Number="singular"<br>POS="ADJECTIVE" id="1.71" offset="554">mincinosul</W> |

Figure 3 POS annotation example

#### 4.4 Classification Based on Sentiment Score

For the first method of classification that we used, namely an averaged sentence score, we needed to find the polarity of the sentence to observe if it expresses a positive or negative opinion.

| clean    | scor_pozitiv | scor_negativ | scor_neutru |
|----------|--------------|--------------|-------------|
| economie | 0.25         | 1.25         | 12.5        |
| creștere | 2.125        | 0.375        | 23.5        |

Figure 4 Sentiment score per word – example

| clean                              | cnt_words | pozitiv | pozitiv_norm | negativ | negativ_norm | neutru | neutru_norm |
|------------------------------------|-----------|---------|--------------|---------|--------------|--------|-------------|
| ['trăim', 'bine']                  | 2         | 20.042  | 10.021       | 1.958   | 0.979        | 46.875 | 23.4375     |
| ['traiasca', 'bine']               | 2         | 19.792  | 9.896        | 1.333   | 0.6665       | 35.75  | 17.875      |
| ['bine', 'zis']                    | 2         | 19.792  | 9.896        | 1.333   | 0.6665       | 35.75  | 17.875      |
| ['buna', 'alegere']                | 2         | 18.875  | 9.4375       | 1.5     | 0.75         | 44     | 22          |
| ['raktrollrahat', 'perje', 'bine'] | 3         | 19.792  | 6.59733      | 1.333   | 0.444333     | 35.75  | 11.9167     |

| clean                                                       | cnt_words | pozitiv | pozitiv_norm | negativ | negativ_norm | neutru | neutru_norm |
|-------------------------------------------------------------|-----------|---------|--------------|---------|--------------|--------|-------------|
| ['nenorociților']                                           | 1         | 1.25    | 1.25         | 6       | 6            | 5.75   | 5.75        |
| ['simțea', 'râu', 'așteptau', 'rezultatul', 'testului']     | 5         | 6.875   | 1.375        | 25.875  | 5.175        | 67     | 13.4        |
| ['uitat', 'râu', 'l', 'făcut', 'aveti', 'tupeul', 'graiti'] | 7         | 16      | 2.28571      | 36.125  | 5.16071      | 356.5  | 50.9286     |

Figure 5 Normalized sentiment score per comment - example

Sentiment classification is viewed as a binary classification (positive or negative), multi-class (negative, neutral, positive) regression or classification. Our approach was to identify the sentimental orientation of the individual words in the comment and then combine them to determine the meaning of the whole comment using a normalized score for each comment, as can be seen in Figure 5.

For each lemma, a positive score, a negative one, and a neutral one were added. Based on this score, a total normalized score was calculated based on the number of words in each comment. Each comment was categorized as *Offensive = Yes* if the condition that the total positive normalized score was higher than the negative was met. The classification result was compared with the manual annotations to evaluate the performance of the model.

#### **4.5 Classification Using Logistic Regression**

The second classification method we tested was supervised Logistic Regression. At this stage we used only the set of manually annotated comments which was divided into a training set (300 observations) and test set (150 observations). In the vectorization process we used two methods: CountVectorizer (converting text to vector with the number of terms / tokens) and TF-IDF (Term Frequency Inverse Document Frequency - compares the frequency of a word with the total number of cases containing that word).

### **5 Discussions**

The two classification methods for identifying offensiveness in social media texts were evaluated using the gold annotated corpus. An overview of the F1-measure obtained with each classification method is presented in the Table 1.

The best model for classifying comments as offensive / non-offensive is the logistic regression model created using the TFIDF Vectorized function, which obtained a correct classification rate for aggressive comments of 87%. This can be easily explained due to the presence of modifiers in texts, which were only limitly considered. Thus, we only use negation as a polarity changer, but the negation only ranges, in our approach, over the next word, and the scope of the negation is not properly computed.

*Table 1 Comparison between the two classifications*

|                                                       | <b>Classification according to the total score obtained</b> | <b>Logistic regression - CountVectorizer</b> | <b>Logistic regression - TFIDF Vectorizer</b> |
|-------------------------------------------------------|-------------------------------------------------------------|----------------------------------------------|-----------------------------------------------|
| <b>Correct classification rate YES (offensive)</b>    | 38,14%                                                      | 73,63%                                       | 87%                                           |
| <b>Correct classification rate NO (non-offensive)</b> | 90,8%                                                       | 38,23%                                       | 26%                                           |
| <b>General correct classification rate</b>            | 58,78%                                                      | 65,27%                                       | 72,91%                                        |

### *Acknowledgements*

This work has been carried out with the financial support of the Romanian Executive Unit for Financing Higher Education, Research, Development and Innovation (UEFISCDI), within the project Computational Representation in Linguistic Big Data Processing (SemRo), project code: PN-III-P1-1.1-PD-2019-0660.

### *References*

- Bhavnani, R., Findley, M. G., & Kuklinski, J. H. (2009). Rumor dynamics in ethnic violence. *Journal of Politics*, 71, 876–892. <https://doi.org/10.1017/S002238160909077X>.
- Castano-Pulgarin, S. A., Suárez-Betancur, N., Tilano Vega, L.M. (2021). Internet, social media and online hate speech. Systematic review. In *Aggression and Violent Behavior*, Volume 58, May–June 2021, 101608, pp. 1-7.
- Dadvar, M., & Niamir, A. (2016). Adopting MaxEnt to Identification of Bullying Incidents in Social Networks. 27th International Workshop on Database and Expert Systems Applications, 186-189.
- Dumitrescu, Ș. D. (2015). RoWordNetLib – The First API For the Romanian WordNet. In *PROCEEDINGS OF THE ROMANIAN ACADEMY*, The Publishing House of the Romanian Academy, Series A, Volume 16, Number 1/2015, pp. 87–94.
- Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). Countering online hate speech. In *Series on internet freedom*, pages 1–73. Unesco: Publishing.
- Henry, S. (2009). *Social deviance*. Cambridge, UK: Polity Press.
- Katakis, I. M., Varlamis, I., Tsatsaronis, G. (2018). PYTHIA: Employing Lexical and Semantic Features for Sentiment Analysis. In *Machine Learning and Knowledge*

- Discovery in Databases Lecture Notes in Computer Science, 2014, Springer Berlin Heidelberg, p. 448-451.
- Rasel, R. I., Sultana, N., Akhter, N., Meesad, P. (2018). Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text mining approach. In Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval, September 2018, pp. 37–41, <https://doi.org/10.1145/3278293.3278303>.
- Van Hee, C., Jacobs, G., Emmery, C., Lefever, E., Verhoeven, B., De Pauw, G., Hoste, V. (2018). Automatic Detection of Cyberbullying in Social Media Text.
- Ventirozos, F. K., Varlamis, I., & Tsatsaronis, G. (2018). Detecting Aggressive Behavior in Discussion Threads Using Text Mining. Lecture Notes in Computer Science, 420-431. <https://www.readcube.com/articles/supplement?doi=10.1038%2Fs41586-019-1494-7&index=1>

# A Survey on Fake News Detection Techniques

RADU CIORA<sup>1,2</sup>, MARIUS CIOCA<sup>1</sup>, and DANIELA GÎFU<sup>3</sup>

<sup>1</sup> *Faculty of Engineering, Lucian Blaga University of Sibiu*

<sup>2</sup> *Faculty of Mechanical and Electrical Engineering,  
University of Petroșani*

<sup>3</sup> *Institute of Computer Science, Romanian Academy – Iasi Branch*

{radu.ciora, marius.cioca}@ulbsibiu.ro,  
daniela.gifu@iit.academiaromana-is.ro

## Abstract

Nowadays, people rely on the information propagated on social networks which allows fast dissemination, therefore becoming viral. In fact, the impact of online media can be seen in all aspects of our lives. Our daily routines are being influenced by it – from buying decisions, to political or religious orientation. Misleading information or fake news are also transmitted by online media. The variety of sources of information, in various formats: text, audio and video are not easy to be verified. To control chaotic spreading, an aggregated news processing engine would be desired to process this vast amount of information, which would be impossible for humans to manage it properly. To find a faster and lighter machine learning technique for identifying what is real from the vast information published online, our main objective is to map the state of art of fake news detection. This paper describes the premises and specification of a text classification system for detection of fake news from online environment, based on a systematic review of recent literature. We conclude that the method for automatic fake news detection is not just one classical machine learning technique, but instead a hybrid approach which should comprise of classic techniques, coordinated by a neural network.

*Key words* — online information, a survey on fake news, misinformation sharing behaviour, binary classification.

## 1 Introduction

Nowadays, revolution is the overused word that describes the impact of internet and online content (e.g., news, user comments, reviews) in the world of information sharing [Talvar *et al.*, 2020]. The success of the online information can be measured, firstly, by the trust degree enjoyed by those who spread it [Delmonte *et al.*, 2013] and only then what is transmitted (the message).

The research question guiding this paper is *how to discern which online information is false or true?* Foremost, the huge amount of news generated online requires a huge process power. Moreover, online information is very colourful,

covering a broad range of topics from various domains, which makes the task of recognition even more difficult.

Basically, to avoid misinformation (fake news), one needs to know how to organize the online information. This in fact tailors an arbitrary process [Kolcun *et al.*, 2014] which needs to be almost instantly. Of particular interest is the ability to detect when online content is untrue or is intended to mislead [Zhang and Ghorbani, 2019].

Therefore, there is no doubt that the evolution of social media has changed the behaviour of information consumers with direct implications in all aspects of their daily lives.

Concerns of online misinformation began with the 2016 presidential election which was considered responsible for influencing the elections' results. According to the research lead by Jennifer Allen [Jennifer Allen *et al.*, 2020], online public disinformation, was not determinant for U.S. elections results, but this does not mean that its impact is marginal. Allcott and Gentzkow remain reluctant to agree with these conclusions [Allcott & Gentzkow, 2017]. They conclude that the impact of social media might be higher than it shows in the numbers given by certain studies.

Moving the attention to our mioritic space, the impact of social media, more specifically Facebook, managed to have a significant importance in changing the final results of the 2014 presidential election in Romania. As the diaspora's main source of information was the internet, with social media playing a major part in this equation [Tasente, 2015]. There is no doubt, that fake news or misleading information about political adversaries are part of every politician's campaign.

The term e-democracy, according to [Benea and Benea, 2015], coins the politician's use of social media for faster interaction with the electorate. The penetration of this type of media is much broader compared to tradition press in general, except for the public television which remains in the lead, at least for now. Nevertheless, social media campaigns have had a significant impact in Romanian electorate, since 2014. It was a tipping point for Romanian politicians to invest in social media marketing and online image.

But it is not just politicians that are the subject of fake news. Stockbrokers or companies' public relations departments are just as active as politicians in both generation and counterfeiting of misleading information. Therefore, the online media seems like a holy grail for fake news dissemination, based on two main reasons. Firstly, the online identity can be easily forged. This means that tracing the source of information can be hard sometimes. Secondly, online credibility can be falsely gained, as fakers can quickly gain "credibility", if the buzzword is interesting enough and engagement of the audience is fed by a hyperactive online presence.

Stock market traders and investors have reasons to spread rumours. By spreading rumours, they influence the stocks' prices, with the aim of gaining

financial benefits by “smartly” trading the influenced stocks. Even though online news is more and more skeptically treated, the trend of digital information consumption is on an ascending trend.

After this preamble, focused on the importance of understanding the place and role of fake news, the rest of the article will focus on the approaches to detecting them, found in the literature. Moreover, we will pursue the best practices in the field and highlight the specifications for an automated news processing application that would independently catalogue news into fake or true.

## **2 Fake News Counterfeiting Methods**

Aiming to reduce the misleading information, as it cannot be eradicated, several approaches were undertaken. This section presents the existing work regarding this matter and proposed approaches. In order to detect the misleading information according to [Albright, 2016], there are two questions that need to be answered: *who is producing the content and what generates the online traffic?*

Also, as suggested in [Zhou and Zafarani, 2018], there is a three-stage process consisting of: a creation stage – in which the article is incepted and written; a second phase is represented by the publishing on the web of the result of the first state; the last stage – is the „putting in stage” or the propagation of the „news” with the aid of online content sharing, comments and reuse of information.

### **2.1 Existing Approaches**

Nowadays, there are several approaches that aim to address the issue of fake news and mis-leading information. The traditional way, by manually verifying every single piece of online content, it is kind of an uphill battle, but still, it is in use today, despite its downside. But, in the era of Artificial Intelligence (AI), the opportunity to filter out fake news is feasible.

The work of fake news detection should not rely on human action, but rather take advantage of the advances in the Natural Language Processing (NLP) and Machine Learning (ML).

According to literature, there are five approaches: (1) linguistic approach; (2) subject-agnostic approach; (3) knowledge-based approach; (4) machine learning approach; (5) hybrid approach.

#### **2.1.1 Linguistic Approach**

There were three different approaches with regard the linguistic approach. The first approach relies on the individual word frequency and using a various Dense neural network, which achieves good prediction rates of the truthfulness of the analysed dataset [Thota *et al.*, 2018].

The second approach, is based on an automated architecture description is presented as a combination of lexical, syntactic, and semantic data, corroborated

with features extracted from the text’s readability property, in the work of [Perez-Rosas *et al.*, 2017].

The last linguistic approach relies on a combination of Naïve Bayes classifier, which is integrated with a support vector machine algorithm. Furthermore, as an enhancement of the proposed system, a latent semantic indexing is suggested. This would enable “deeper understanding” of the content by machines [Stahl, 2018].

### **2.1.2 Subject-Agnostic Approach**

In this approach, the content of the analysed material is ignored and the subject of the investigation becomes its surrounding content which proved useful according to de Beer and Matthee [2021]. More specifically, the subject takes comprises of the number of advertisements present on the analysed page, the length of the headline and its eye catchiness, various patterns extracted from mainstream media sources which induce subjective behaviour and the effective assumption of the content by an author.

Moreover in [Castelo *et al.*, 2019] HTML markup information is also included in their approach, which proves to be effective in improving the fake content detection algorithms’ performance.

Along the above lines, is the work of [Chen and Freire, 2020] which based their approach on the observation that while the news topic may change day by day, the layout and writing style of a website do not change as frequent, nor does the writing style of an author, thus they focused on the writing style and page layout style.

### **2.1.3 Knowledge-Based Approach**

In this approach, the process is called fact-checking – that means that an individual manually checks the value of truth of the article which is under assessment against a knowledgebase extracted from a trustable source. The second approach uses an automated computer program to do a similar workload. The former is called manual fact-checking, while the latter is coined automated fact-checking.

In the manual fact-checking approach the data is check by domain-experts – the so-called fact-checkers, which can provide a correct coining of the pieces of information as either true or false. There are several websites like PolitiFact, FactCheck or Snopes that enable users to verify online content.

A recent trend in this approach can be seen on Fiskkit – a web site that promotes crowd-based fact-checking services. Here user can upload documents (articles) and based on ratings and tagging, a multi-class categorisation takes place, which enables a better study of the patterns encountered over the new articles [Zhou and Zafarani, 2018].

Automated fact-checking overcomes the issues of the manual fact-checking approach, by means of a computer software which comprises of information retrieval, NLP algorithms, and graph theories [Cohen *et al.*, 2011].

#### 2.1.4 *Machine Learning Approach*

Automated learning approach makes use of machine learning algorithms. A common approach is the use of crowdsourcing, which provides the necessary dataset for this class of algorithms to perform well, as they need two distinct sets of data to function properly. The first dataset is the training set, from which the algorithms learn how to distinguish fake vs real news. The second is the test dataset against the algorithm is tested to check its performance [Zhou and Zafarani, 2018].

Another extensive study was undertaken by [Khan *et al.*, 2019], which concluded that a traditional machine learning Multinomial Naive Bayes Model can have good results in detection of fake news, but also deep learning techniques look promising, especially on larger dataset (of 5000+ news). The paper suggests that LSTM (Long short-term memory)-based models are dataset dependent, but their advantage is that they can overcome overfitting.

The work led by Sharma, analysis four machine learning algorithms: a Naive Bayes classifier, the random forest deep learning algorithm, the logistic regression algorithm performance and the classification results of passive aggressive algorithm. In their implementation the best performing algorithm was the logistic regression [Sharma *et al.*, 2020].

#### 2.1.5 *Hybrid Approach*

In the hybrid approach, the focus is on three main elements of an article: (1) the text of the article; (2) The feedback received by the articles; (3) the source or author of the article.

Such a hybrid model is described in [Ruchansky *et al.*, 2017], where a model called CSI (capture, score and integrate) was developed. The model comprises of three parts: one that uses a neural network to map the user activity over time; a second part determines the quality of the source and the last part of the algorithm corroborates the information of the former parts and presents the output as true or false.

### 3 *Conclusions*

The briefly mentioned approaches presented above, reveal a simple fact. The linguistic approaches, although they might seem at first sight a bit trivial, when combined with machine learning algorithms, they prove to perform very good.

Knowledge-based approaches, because of the large number of articles that are generated over time, have very little applicability in practice. The reason for that is clear: it is impossible to manual fact-checking every single article that is published online.

Regarding the machine learning approaches, even though the cited papers' results of previous research conclude different best performing algorithms, the

general idea is that the bigger the dataset, the better the performance of the algorithms.

Hybrid approaches tend to grasp the big picture better related to this issue and so, we conclude that they seem the most appropriate approaches, which can tackle the variety of news contents that are out there in the online environment. Moreover, the use of HTML markup analyses also can have a positive impact in improving the detection accuracy.

#### 4 *Future Work*

The source or the author of news content is an important factor that can play a key role in news classification. The role of argumentation features in news classification depends on both the source and the author of a message. To build an online identity of the author can be a good starting point, from which can be built upon.

Another future direction of research would be the involvement of sentiment analysis, to improve the accuracy of fake news detector, as suggested in [Gifu and Cioca, 2014] and [Bhutani *et al.*, 2019].

Based on the investigations presented in this paper, a hybrid approach seems to be the most suitable way in addressing the fake news issues, especially when we deal with various types of news. Moreover, by enriching the hybrid approach with a HTML mark-up analysis tool would certainly improve the prediction accuracy of such a model.

#### *References*

- Albright, J. (2016, November 18). The #Election2016 Micro-Propaganda Machine. Retrieved from Medium – Where good ideas find you.: <https://d1gi.medium.com/the-election2016-micro-propaganda-machine-383449cc1fba#.14h9hafbd>
- Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, 31(2), 211–236. doi:10.1257/jep.31.2.211
- de Beer D., Matthee M. (2021) Approaches to Identify Fake News: A Systematic Literature Review. In: Antipova T. (eds) *Integrated Science in Digital Age 2020. ICIS 2020. Lecture Notes in Networks and Systems*, vol 136. Springer, Cham. [https://doi.org/10.1007/978-3-030-49264-9\\_2](https://doi.org/10.1007/978-3-030-49264-9_2)
- Benea, M. C., & Benea, O. D. (2015). The Use of Social Media in Romanian Political Marketing. *Review of the Air Force Academy*, 2(29), 103-106.
- Bhutani, B., Rastogi, N., Sehgal, P., & Purwar, A. (2019). Fake News Detection Using Sentiment Analysis. 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-5). Noida: India. doi:10.1109/IC3.2019.8844880
- Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., Freire, J. (2019). A Topic Agnostic Approach for Identifying Fake News Pages. In: *Companion Proceedings of the 2019, World Wide Web Conference on - WWW 2019*, pp. 975–980 (2019). <https://doi.org/10.1145/3308560.3316739>.

- Chen, Zhouhan & Freire, Juliana. (2020). Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds. 584-592. 10.1145/3366424.3385772.
- Cohen, S., Hamilton, J., & Turner, F. (2011). Computational Journalism. *ACM Communications*, 54(10), 66-71.
- Delmonte, R., Tripodi, R., Gifu, D. (2013). Opinion and Factivity Analysis of Italian Political Discourse. In: *Proceedings of the 4th edition of the Italian Information Retrieval Workshop (IIR 2013)*, Pisa, Italy, CEUR-WS on-line proceedings series, pp. 88-99.
- Gifu, D., & Cioca, M. (2014). Detecting Emotions in Comments on Forums. *International Journal of Computers, Communications & Control*, 9(6), 694-702. doi:10.15837/ijccc.2014.6.1474
- Jennifer Allen, B. H. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, 6(14). doi:10.1126/sciadv.aay3539
- Khan, J., Khondaker, M., Iqbal, A., & Afroz, S. (2019). A Benchmark Study on Machine Learning Methods for Fake News Detection. *arXiv*, abs/1905.04749.
- Kolcun, M., Teodorescu, B., Gifu, D. (2014). Elements of Organizational Communication. In: *International Letters of Social and Humanistic Sciences*, Vol. 24, pp. 66-73
- Perez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2017). Automatic Detection of Fake News. *arXiv*, abs/1708.07104.
- Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: a hybrid deep model for fake news detection. *ACM on Conference on Information and Knowledge Management*, (pp. 797-506). doi:10.1145/3132847.3132877
- Sharma, U., Saran, S., & Patil, S. (2020). Fake News Detection Using Machine Learning Algorithms. *International Journal of Creative Research Thoughts*, 8(6).
- Stahl, K. (2018). Fake news detection in social media. Turlock, California, USA.
- Talwar, S., Dhir, A, Singh, D., Singh Virk, G., Salo, J. (2020). Sharing of fake news on social media: Application of the honeycomb framework and the third-person effect hypothesis. *Journal of Retailing and Consumer Services*, Vol. 57, pp. 102-197 – DOI 10.1016/j.jretconser.2020.102197
- Tasente, T. (2015). The electoral campaign through Social Media. Case Study – 2014 Presidential elections in Romania. *Sfera Politicii*, 183(1), pp. 92-104.
- Thota, A., Tilak, P., Ahluwalia, S., & Lohia, N. (2018). Fake News Detection: A Deep Learning Approach. *SMU Data Science Review*, 1(3).
- Zhang, X., Ghorbani, A.A. (2019). An Overview of Online Fake News: Characterization, Detection, and Discussion. *Information Processing and Management* 57(2) - DOI: 10.1016/j.ipm.2019.03.004
- Zhou, X., & Zafarani, R. (2018). Fake news: A survey of research, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2.



## INDEX OF AUTHORS

Avram, Andrei-Marius, 7, 165

Barbu Mititelu, Verginica, 7, 165

Bejinariu, Silviu-Ioan, 75

Berindei, Cosmina-Maria, 37

Bibiri, Anca-Diana, 181

Buhnilă, Ioana, 139

Cioca, Marius, 207

Ciora, Radu, 207

Colhon, Mihaela, 181

Cristea, Dan, 59

Curea, Eric, 7, 165

Donu, Bogdan-Andrei, 199

Feraru, Monica, 155

Floarea, Carmen Irina, 75

Gîfu, Daniela, 207

Groza, Adrian, 119

Haja, Gabriela, 49

Iftene, Florin, 75

Ion, Radu, 165

Irimia, Elena, 7, 165

Mitrofan, Maria, 7, 165

Mocanu, Mihaela, 181

Moruz, Mihai Alex, 93

Nevaci, Manuela, 75

Nicolae, Dragoș Constantin, 105

Niculescu, Oana, 21

Onofrei, Mihaela, 59, 199

Pădurariu, Cristian, 59

Păiș, Vasile, 7, 165

Rebeja, Petru, 59

Scutelnicu, Andrei, 59

Szomiu, Roxana, 119

Tamba, Elena Isabelle, 49

Trandabăț, Diana, 199

Tușiș, Dan, 105, 165

Ungureanu, Mădălina, 93

Zbancioc, Marius, 155