Interaction Design and Architecture(s) Journal - IxD&A, N.60, 2024, pp. 7 - 26
DOI: 10.55612/s-5002-060-001sp

Natural Language Processing Tools for Romanian – Going Beyond a Low-Resource Language

Melania Nitu[1], Mihai Dascalu[1,2]

[1] University Politehnica of Bucharest, Faculty of Automated Control and Computers,
Splaiul Independenței 313, 060042, Bucharest, Romania
[2] Academy of Romanian Scientists, Str. Ilfov, Nr.3, 050044, Bucharest, Romania
melania.nitu@yahoo.com, mihai.dascalu@upb.ro

Abstract. Advances in Natural Language Processing bring innovative instruments to the educational field to improve the quality of the didactic process by addressing challenges like language barriers and creating personalized learning experiences. Most research in the domain is dedicated to high-resource languages, such as English, while languages with limited coverage, like Romanian, are still underrepresented in the field. Operating on low-resource languages is essential to ensure equitable access to educational opportunities and to preserve linguistic diversity. Through continuous investments in developing Romanian educational instruments, we are rapidly going beyond a low-resource language. This paper presents recent educational instruments and frameworks dedicated to Romanian, leveraging state-of-the-art NLP techniques, such as building advanced Romanian language models and benchmarks encompassing tools for language learning, text comprehension, question answering, automatic essay scoring, and information retrieval. The methods and insights gained are transferable to other low-resource languages, emphasizing methodological adaptability, collaborative frameworks, and technology transfer to address similar challenges in diverse linguistic contexts. Two use cases are presented, focusing on assessing student performance in Moodle courses and extracting main ideas from students' feedback. These practical applications in Romanian academic settings serve as examples for enhancing educational practices in other less-resourced languages.

Keywords: Natural Language Processing, Educational Frameworks, Romanian Language Models, Transformer Architecture.

1 Introduction

Communication and activities in educational environments occur through speech and text, making Natural Language Processing (NLP) instruments excellent candidates for enhancing learning. While the spotlight often falls on high-resource languages, the significance of NLP instruments for low-resource languages cannot be underestimated. As such, this paper focuses on NLP applications in the context of Romanian, a low-resource language with specific challenges. Nonetheless, the insights provided by this study extend beyond the geographical boundaries or linguistic specificity, representing a blueprint for researchers working on other low-resource languages by showcasing adaptable methodologies and strategies that surpass the limitations of insufficient data. NLP contributes to education by addressing various problems and challenges and by optimizing existing learning instruments. As such, NLP is widely used in academic settings to enhance teaching and learning activities while providing assistance in multiple contexts – for example, language learning [1, 2, 3, 4, 5], text summarization [6, 7, 8] and paraphrasing [9, 10, 11], question answering [9, 12, 13], automated essay scoring [14, 15, 16], personalized learning and feedback [17, 18], intelligent tutoring [19], and information retrieval [20]. All previous contexts are presented briefly from a global perspective in the following paragraphs, whereas details for Romanian are introduced in subsequent sections.

In language learning [1, 2], NLP improves the effectiveness of linguistic instruments while identifying and correcting grammatical, syntactic, and spelling errors [3, 5, 21, 22]. NLP is also used to identify text patterns and the presence of discourse elements such as topic, coherence, or coreference structure [23, 24, 25, 26]. Moreover, NLP tools can provide formative feedback on student writing and assist in vocabulary acquisition, discourse cohesion, pronunciation improvement, and speech recognition, enhancing the academic performance of students [27, 28, 29, 30].

Since text comprehension is a difficult task, summarization [31, 32] can reduce text length and the underlying complexity by highlighting the main concepts and extracting the most relevant information. To further enhance clarity and reduce difficulty, paraphrasing can also be employed in conjunction with text summarization to ensure an accessible reading level.

While considering intelligent assessment systems, question-answering models [9, 12] were developed to generate customized questions and answers from learning materials. These models can tailor custom questions depending on students' learning pace or reading level while considering students learning data.

A valuable educational application of NLP in academic settings is automated essay scoring [16, 33, 34, 35], which facilitates the grading of essays and written assignments. These systems can analyze factors like grammar, vocabulary, and coherence to provide quick and accurate feedback to students and teachers.

Personalized learning is a meaningful instrument, enabling tutors to provide customized learning and tailor educational content to students' individual needs based on the analysis of learning patterns. Intelligent tutoring systems [19, 36, 37] emerged to provide personalized guidance and feedback based on learning progress and performance [18]. In addition, information retrieval proved its great efficiency in education, helping retrieve, analyze, and integrate information from large volumes of textual data from various sources.

Overall, NLP enhances learning instruments and improves the quality of the educational process by tackling several challenges, such as language barriers, time constraints, and matching individual learning needs, while increasing student engagement and participation.

The structure of this paper comprises several sections that systematically present recent educational instruments and frameworks dedicated to Romanian while leveraging state-of-the-art NLP techniques. The second section contextualizes the study by introducing the motivation and relevance of the current background. It is followed by a section summarizing the latest publicly available datasets for Romanian. The fourth section details the Romanian language models, while the next two sections present cutting-edge general and educational Romanian NLP frameworks. The paper also highlights two research studies with applications in academic settings, namely assessing students' performance in Moodle courses and extracting and clustering main ideas from students' feedback. Finally, the last section of the paper presents the conclusions drawn from this comprehensive review.

2 Relevance and Motivation

The scope of this study focuses on NLP methodologies in education and extends beyond the Romanian use case, being relevant to researchers, linguists, and NLP enthusiasts globally. While the immediate focus is on Romanian resources, the methodologies, insights, and frameworks explored in this paper can be transferred and applied to a multitude of languages. The challenges encountered in dealing with a low-resource language like Romanian often mirror those faced by other languages with limited data. Thus, this study's findings have the potential to serve as a bridge, connecting language-specific insights to universal NLP practices. This paper offers a broader takeaway for readers unfamiliar with the intricacies of Romanian. The challenges discussed, ranging from low data availability to domain adaptation, are inherent to numerous languages across the globe. The universal message lies in the adaptability and transferability of NLP techniques, highlighting the potential for collaboration and knowledge exchange across linguistic boundaries.

Although research in NLP has rapidly advanced in recent years with the introduction of the Transformer [38] and with numerous studies and applications in the educational field, only a few apply to the low-resource languages. Introduced as a paradigm shift from sequential processing in Recurrent Neural Networks, the Transformer model reformed automated language understanding and generation through its attention mechanisms and the simultaneous processing of all words in a text sequence, thus enabling models to capture long-range dependencies and contextual relationships in an efficient way. This architecture's capability to understand the context and generate coherent responses led to more interactive and dynamic learning experiences. As a result, the Transformer architecture significantly contributed to modernizing education, making it more engaging, adaptive, and effective.

By investing in NLP educational tools tailored to low-resource languages, we empower communities to participate actively in enabling information access and facilitating communication in native languages. Despite having around 24 million native speakers, the Romanian language was identified not long ago as a low-resource language with only a few freely available digital resources and NLP instruments [39]. Due to continuous investment and the growing interest of the scientific community in the research field, Romanian is rapidly evolving towards a mid-range resource language. According to the European Language Grid[1], Romanian has 642 resources nowadays, compared to high-resource languages such as English (6,215), Spanish (2,561), or German (2,379).

[1] https://live.european-language-grid.eu/catalogue/, last accessed 07/10/2023.

In this context, this paper aims to provide a comprehensive review of the current state-of-the-art NLP instruments for Romanian, highlighting the interdisciplinary nature of NLP and inviting experts and practitioners from diverse fields to collaborate and innovate collectively to improve educational systems. The current study highlights the importance of low-resource languages, such as Romanian, as valuable resources for NLP research, particularly regarding their linguistic features and cultural significance. While transitioning to a mid-range resource language, Romanian presents unique challenges and opportunities for NLP development, and this paper sheds light on the latest advancements in this field. The review study provides valuable insights into the latest developments in NLP instruments for Romanian, which can benefit researchers, developers, and tutors working with this language. The paper concludes by emphasizing the need for continued research and development of NLP instruments for Romanian, as the language is moving beyond its low-resource status and gaining more recognition in the global NLP community.

3 Romanian Corpora

Consistent efforts are invested in developing linguistic corpora to create reference language resources for Romanian. For this purpose, academics have built several publicly available and open-sourced datasets suitable for different NLP tasks. Table 1 incorporates some of the most popular text and bi-modal corpora for the Romanian language, while Table 2 focuses on Romanian speech datasets.

Table 1. General Romanian Corpora.

Dataset name	Description
RoWordNet	Romanian WordNet [47] represents a semantic network based on synonym sets split by part-of-speech (PoS) tags (i.e., nouns, verbs, adverbs, and adjectives). The lexicalized ontology provides semantic relations between Romanian words, such as hypernymy, meronymy, or antonymy.
ROMBAC	The Romanian Balanced Annotated Corpus [48] incorporates 41 million words while providing morphosyntactic information. It contains equal shares of text from five domains: legal, journalism, fiction, medicine, and biography of Romanian personalities.
RONEC	Romanian Named Entity Corpus (RONEC) v2.0 [49] contains over 12,000 sentences, with more than 0,5 million tokens and 80,000 distinctly annotated entities belonging to 15 different classes.
RoTex	RoTex [50] is a Romanian plain text corpora extracted from various online sources containing writings from different genres: political, bibliography, literature, newspapers, juridical, and Wikipedia dumps. It sums up over 1 billion tokens, having 62,22% DEX[2] coverage.
RONACC	RONACC [51], also called RoGEC[3], is the first Romanian corpus targeting Grammatical Error Correction with 10,000 pairs of sentences.

[2] https://github.com/dexonline/dexonline, last accessed 07/10/2023.
[3] https://github.com/teodor-cotet/RoGEC, last accessed 07/10/2023.

Dataset name	Description
CoRoLa	The Contemporary Romanian Language (CoRoLa)[4] [52] is a bi-modal corpus (text and speech) containing over 1 billion tokens, 70 scientific domains, and 300 hours of recordings.
MOROCO	MOROCO (Moldavian and Romanian Dialectal Corpus) [53] contains over 33,000 text samples, with over 10 million tokens collected from the news.
CoRoSeOf	CoRoSeOf [54] is a manually annotated dataset for Romanian sexist and offensive language collected from approx. 40,000 tweets.
ROFF	ROFF [55] is a Romanian Twitter dataset for offensive language, containing 5,000 micro-blogging annotated posts.
Ro-Offense	RO-Offense [56] is the largest Romanian language dataset for offensive language, consisting of over 12,000 online manually annotated comments extracted from sports websites.
News-RO-Offense	An offensive language dataset representing a collection of 4,052 manually annotated comments gathered from local Romanian news websites [57].
RoCo-News	RoCo [58] is a journalistic corpus counting 7,1 million words and 231,626 distinct tokens. The corpus is lemmatized and annotated with PoS tags.
AlephNews	Dataset for Romanian text summarization[5] built by crawling 48,862 articles (news and summary text) from https://alephnews.ro/ until July 2022.
XquAD-ro	Cross-lingual Question Answering Dataset (XquAD) [59] consists of 240 paragraphs and over 1,000 QA pairs translated into 11 languages. The Romanian component of XquAD was introduced with LiRo [60].
RoITD	RoITD is a Romanian IT Question Answering Dataset [61] with over 9,500 QA pairs extracted from 5,000 Romanian Wikipedia articles describing IT products.
RO-STS	RO-STS is the Romanian translation of the Semantic Textual Similarity (STS)[6] dataset and was introduced with LiRo [60]. The corpus contains 8,628 sentence pairs with their similarity scores.
ROST	Romanian Stories and other Texts (ROST) [62] contains 400 texts written by 10 authors, representing short stories, literature, novels, articles, or sketches.
ROGER	The bilingual corpus of Romanian Academic Genres (ROGER)[7] [63] is a learner corpus of bilingual student academic writing in Romanian and English. The corpus represents a collection of 2,050 texts from 8 disciplines, with over 20 million tokens.
Oscar	Oscar (Open Super-large Crawled Aggregated coRpus) [64] is a multilingual dataset obtained via filtering the common crawl corpus. It encompasses 151 languages. The Romanian sub-corpus contains over 4,5 million documents, encapsulating approximately 5 billion tokens.

Besides general corpora, the research community has also created specialized datasets for different domains, such as legal or medical ones. For the legal domain, we mention MARCELL [40], which contains 7 monolingual sub-corpora (Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak, and Slovenian). The Romanian version contains over 163,000 documents representing national legislation texts originating between 1881 and 2021. LegalNERo [41] is a Romanian Named Entity Recognition (NER) corpus for the legal domain containing a subset of 370 documents from MARCELL-Ro. A sub-section of the juridical field is represented by

[4] https://corola.racai.ro/, last accessed 07/10/2023.
[5] https://huggingface.co/datasets/readerbench/AlephNews, last accessed 07/10/2023.
[6] https://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark, last accessed 07/10/2023.
[7] https://roger-corpus.org/, last accessed 07/10/2023.

parliamentary debates incorporated into ParlaMint-Ro [42], a corpus with 1,832 transcribed sessions and 552,104 processed speeches collected over 20 years between 2000 and 2020.

Table 2. Romanian Speech Datasets.

Dataset name	Description
Echo	Echo[8] is a mixture of genres with more than 100,000 recordings from 280+ speakers, with a total duration of over 230 hours and more than 4,300 unique transcripts having more than 1,8 million tokens.
RSC	Romanian Read-Speech Corpus (RSC)[9] consists of 100 hours and 136,120 audio files collected from 164 Romanian native speakers.
Swara	Swara[10] is a Romanian read speech dataset containing over 21 hours of recordings from 17 different speakers.
CoBiLiRo	CoBiLiRo[11] [65] is an annotated bi-modal corpus summarizing a selection of Romanian speech data with more than 520 hours of transcriptions.

The medical domain is also strongly represented among language resources. The Biomedical Corpus for the Romanian Language (BioRo) [43] comprises over 322,000 tokens in 18,000 sentences from 7 medical fields, namely neurology, diabetes, endocrinology, cardiology, oncology, nephrology, and alternative medicine, extracted from online resources. Another example is MoNERo [44], a biomedical corpus of contemporary Romanian, morphologically, syntactic, and named entity annotated, with over 154,000 tokens and 23,000 entity annotations belonging to 4 semantic groups. We further reference SiMoNERo [45, 46], a medical corpus containing more than 163,000 tokens from over 5,400 sentences and 15,490 named entities. The text was collected from books, journal articles, and blog posts from 3 medical fields: cardiology, diabetes, and endocrinology.

Given the numerous datasets, we observe a recent noticeable advancement in Romanian corpora resources, favoring the context for NLP development.

4 Romanian Language Models

Over the past few years, Transformer-based language models have become widely used in NLP research, replacing traditional word embeddings with pre-trained models capable of understanding contextual information across entire sequences. The Transformer architecture evolved into a key component of numerous architectures, including BERT (Bidirectional Encoder Representations from Transformers) [66], GPT (Generative Pre-Trained Transformer) [67, 68], or XLNet (autoregressive method to learn bidirectional contexts) [69]. However, most of these models are designed for high-resource languages like English, leaving languages with fewer resources, such as

[8] https://echo.readerbench.com/, last accessed 07/10/2023.
[9] https://speed.pub.ro/downloads/speech-datasets/, last accessed 07/10/2023.
[10] https://speech.utcluj.ro/swarasc/, last accessed 07/10/2023.
[11] http://cobiliro.info.uaic.ro/, last accessed 07/10/2023.

Romanian, at a disadvantage. To bridge this gap, dedicated Romanian-specific language models based on BERT have emerged.

One such model is RoBERT [70], trained on Romanian text from sources like Wikipedia, Oscar [64], and RoTex [50], with a vocabulary size of 38,000 tokens. During training, RoBERT followed the original BERT approach, which involves two key tasks: masked language modeling (MLM) and next sentence prediction (NSP). The model predicts masked tokens within a given sentence in the MLM task, helping it learn the contextual relationships between words. The NSP task determines whether two sentences are consecutive or randomly sampled from the dataset, supporting the model in understanding the coherence of text sequences. To handle the complexity of Romanian text, RoBERT considers diacritics and utilizes WordPiece tokenization to break down words into smaller sub-word units, enabling it to capture detailed linguistic features. The model is available in three configurations: RoBERT-small, RoBERT-base, and RoBERT-large, each differing in parameters, layers, hidden layers, and attention heads. Despite their varying complexities, all configurations have achieved competitive performance in various Romanian-specific NLP tasks, including sentiment analysis, cross-dialect topic identification, and automated diacritics restoration.

Dumitrescu et al. [71] developed a similar version of Romanian BERT, named BERT-base-ro. The model was pre-trained on Opus [72], Oscar [64], and Wikipedia datasets, resulting in a vocabulary of 50,000 words. The same BERT methodology was applied for training. Unlike RoBERT, this model was trained only on one supervised task, namely MLM. Two versions of the model were released: cased and uncased. For tokenization, the model uses Byte-Pair Encoding (BPE), which splits words into smaller sub-words units. Although BERT-base-ro performed better than multilingual BERT (mBERT), it scored lower than RoBERT-large on most comparisons conducted by Masala et al. [70]. The evaluation included simple universal dependencies, joint universal dependencies, and named entity recognition with label prediction.

With the introduction of RoGPT2 [73], state-of-the-art performance was achieved for Romanian text generation by leveraging a Romanian version of GPT2 [74]. RoGPT2 was trained using the largest available corpus for Romanian and was evaluated against 6 tasks from the LiRo benchmark [60]. These tasks included text categorization and dialect classification, sentiment analysis, semantic textual similarity, machine translation, QA with zero-shot cross-lingual learning, and language modeling. RoGPT2 achieved superior performance compared to other BERT-based models for Romanian, such as RoBERT or BERT-ro-base, across most tasks, except for zero-shot cross-lingual learning. It also showed competitive results in grammar error correction (RoGEC), utilizing the RONACC corpus to generate grammatically correct text. RoGPT2 was released in three versions: base (124M parameters), medium (354M parameters), and large (774M parameters).

RoSummary [31] is a language model designed for creating summaries. The model is based on the RoGPT2 architecture, trained to predict the next token using the previous sequence. To control the output, four tokens were used to indicate the characteristics of generated text: NoSentences (indicating the number of sentences that the summary should have), NoWords (specifying the number of words generated in the summary), RatioTokens (determining the proportion of words in the summary compared to the input text), and LexOverlap (measuring the ratio of 4-grams in the summary that appear in the reference text). The model produced grammatically correct summaries and was evaluated using ROUGE and BERTScore. Three versions of the model were tested and released, each with a different context size: base (12 layers, batch size of 128), medium (24 layers, batch size of 24), and large (36 layers, batch size of 16). Among these, the medium version yielded the best results, achieving a ROUGE score of 34.67% and a BERTScore of 74.34%. Additionally, the consideration of control tokens resulted in a slight improvement in BERTScore by up to 2%. Notably, using only one control token tended to produce higher scores overall.

A newer model, GPT-NEO-RO[12], is based on the GPT-3 architecture and is one of the largest language models available for Romanian, with 780M parameters. Its design consists of multiple transformer blocks linked by multi-head self-attention mechanisms. This setup helps the model understand connections between different parts of a sentence, which aids in creating understandable and fluent text. GPT-NEO-RO was trained on a massive 40GB collection of Romanian text from diverse sources like Oscar, Wikipedia, and Romanian literature. It was fine-tuned on various NLP tasks, such as language modeling, text classification, and sentiment analysis [79].

While dedicated language models generally outperform multilingual models, few multilingual alternatives exhibit good results. One such model is mBart [80], which is used for Machine Translation tasks. It was trained on large-scale datasets and has shown improvements when pre-trained on a Romanian-English corpus. Additionally, there are other multilingual models like mT5 [82] and Flan-T5 [83]. mT5 is trained in 101 languages and achieved good performance on various tasks. Flan-T5 is an enhanced version of T5, available in different sizes, and has shown promising improvements in performance across different tasks.

Romanian versions of T5 and Flan-T5 were recently published by the research community[13] and fine-tuned for specific tasks, such as paraphrasing. Flan-T5-paraphrase-ro is built on the T5 architecture and was fine-tuned for the paraphrasing task. The model generates different versions of the same sentence while preserving its meaning. It was pre-trained on 60,000 Romanian paraphrasing documents[14], and it was released in three sizes: small (77M parameters), base (220M parameters), and large (783M parameters).

When comparing dedicated language models like RoBERT, RoGPT2, RoSummary, and GPT-Neo-Ro with multilingual models such as mBART and Flan-T5, their performance varies based on the availability of linguistic resources. Dedicated models significantly improve in low-resource settings, better capturing language complexity and generating more coherent text. Dedicated models maintain their superiority as resource availability increases from low to medium or high, producing more accurate and relevant text. However, multilingual models perform well across multiple languages but may lack linguistic nuance in low-resource settings due to their broad focus.

The language resource level, either low, medium, or high, impacts the performance of language models significantly. Dedicated models excel in low-resource languages by offering language-specific capabilities, while multilingual models are stronger in medium to high-resource settings. The performance of multilingual models relies

[12] https://huggingface.co/dumitrescustefan/gpt-neo-romanian-780m, last accessed 7/16/2023.
[13] https://huggingface.co/BlackKakapo, last accessed 07/16/2023.
[14] https://huggingface.co/datasets/BlackKakapo/paraphrase-ro, last accessed 07/16/2023.

heavily on the size and quality of the corpus used for training. Unlike dedicated language models, which undergo extensive pre-training on large, language-specific text datasets, multilingual models often cover a wide array of languages without detailed pre-training for each specific language. This discrepancy in complexity between multilingual and dedicated models can originate from various factors, including the volume of training data, differences in model architecture and size, and unique linguistic characteristics of languages. This highlights the importance of dedicated solutions for low-resource languages and underscores the adaptability of multilingual models for languages with more resources.

5 General NLP Tools Dedicated to Romanian

Recent studies like the one by Ranathunga et al. [84] assessed the research levels, linguistic gaps, and data availability for various languages, offering suggestions to improve low-resource languages. Khan [85] compared the performance of different NLP tools for high-resource languages, like English. Hershcovich et al. [86] identified challenges in cross-cultural NLP that can be addressed via specialized instruments. Additionally, Aji et al. [87] extensively discussed challenges faced by underrepresented languages and dialects in Indonesia.

Comparing NLP tools dedicated to Romanian with tools designed for other languages shows how linguistic differences, resource availability, and tool performance interact. The development of language resources and the appearance of dedicated language models led to the emergence of Romanian NLP platforms.

Dumitrescu et al. [60] introduced LiRo[15] (**Li**mba **Ro**mână = Romanian Language), a benchmark comprising 8 datasets and 10 tasks covering various aspects like text categorization, named entity recognition, machine translation, language modeling, PoS tagging, dependency parsing, question answering, sentiment analysis, semantic textual similarity, and gender debiasing of language embeddings. The platform aligns with international NLP benchmarks, facilitating cross-lingual studies.

Another significant resource is RELATE[16] web platform [88], which integrates diverse language technologies, tools, and datasets for Romanian, offering advanced text and speech processing capabilities. It supports tasks like text annotation, syntactic analysis, and dataset development, with a search feature for the CoRoLa corpus.

Additionally, spaCy[17], a popular open-source library, provides Romanian-trained pipelines in three versions: small (12 MB), medium (40 MB), and large (542MB). These pipelines, trained on news and media data, incorporate advanced features like convolutional neural networks for tasks such as part-of-speech tagging and named entity recognition.

[15] https://www.airomania.eu/projects/liro, last accessed 07/10/2023.

[16] https://relate.racai.ro/, last accessed 07/10/2023.

[17] https://spacy.io/models/ro, last accessed 07/10/2023.

6 Educational NLP Instruments Dedicated to Romanian

Concurrently with the latest research progress in the NLP, intelligent systems with educational purposes were developed for Romanian using the previously presented language models. ReaderBench[18] is a multilingual framework supporting 8 languages, i.e., English, French, Romanian, Spanish, German, Russian, Italian, and Dutch. ReaderBench encompasses various text analysis modules, including textual complexity evaluation and cohesion-based collaboration assessment [89]. The textual complexity indices[19] cover multiple levels, including surface and lexicon (e.g., statistics on words/character/n-gram counts, punctuation marks, word entropy), morphology and syntax (e.g., PoS occurrences, syntactic dependencies, parse tree depth), semantics (e.g., cohesion, coreferences, document cohesion flow), discourse structure (e.g., presence of specific connectors) and word complexity (e.g., characters, flectional forms, syllables, Wordnet statistics, and various word lists). The indices are applied at different granularity levels (i.e., document, paragraph, sentence, or word) and using different aggregation functions (i.e., mean, standard deviation, and maximum). ReaderBench also includes an automated pipeline built on top of these indices that targets automated scoring of texts written in Romanian.

ReaderBench is designed to be scalable and easily extensible, promoting both individual and collaborative learning. These features are highlighted in a recent study by Dascalu et al. [90] that presented an enhanced version of the framework used as a Moodle plug-in to predict students' performance. The analysis targeted timeframes before and during the COVID-19 pandemic. Students' interactions were compared using data from an Algorithm Design course collected from two successive years: before and during the pandemic. The corpus consisted of a collection of forum posts with anonymized usernames, timestamps, and reply-to links from each discussion thread and the online activity extracted from click-stream log data in two consecutive academic years (i.e., 2018-2019 before the COVID-19 pandemic and 2019-2020 during the outbreak). Cohesion Network Analysis (CNA) [91] is the building block of the experiment. A CNA cohesion graph is created as a proxy for the semantic content of discourse while assessing the active engagement of students within the discourse. The considered features for predicting student grades are illustrated in Fig. 1. Starting from the previous graph, specific CNA indices (e.g., in-degree, out-degree, and other centrality measures) are used to evaluate student interactions. The CNA initiation indices refer to the community's activity following the new discussion thread initiation. The indices refer to new threads (conversation threads initiated by a given participant), the overall score (sum of contribution scores from initiated discussion threads by a given participant), and the average length (average count of distributions per initiated discussion thread). Lastly, features from time series analysis are generated in a longitudinal analysis that models evolution across time; these features are divided into two categories: features applied on specific timeframe CNA indices (e.g., local extreme points and slope) and global longitudinal analysis (e.g., M&SD recurrence).

[18] https://readerbench.com/, last accessed 07/10/2023.

[19] https://github.com/readerbench/ReaderBench/wiki/Textual-Complexity-Indices, last accessed 07/10/2023.

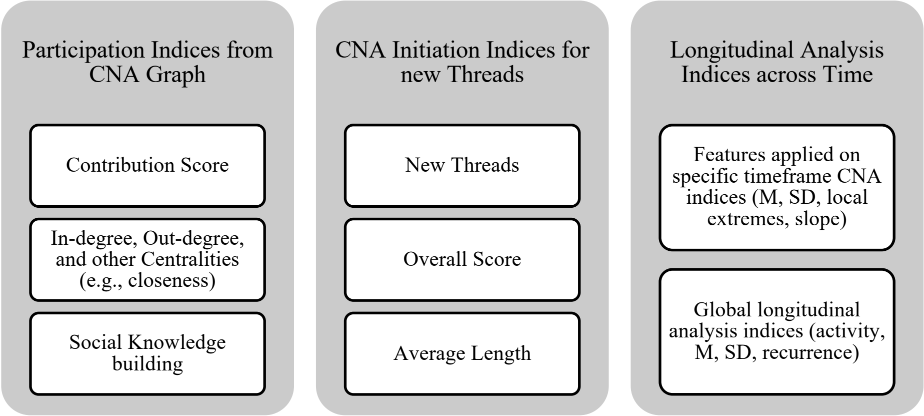
Participation Indices from CNA Graph
Contribution Score
In-degree, Out-degree, and other Centralities (e.g., closeness)
Social Knowledge building
CNA Initiation Indices for new Threads
New Threads
Overall Score
Average Length
Longitudinal Analysis Indices across Time
Features applied on specific timeframe CNA indices (M, SD, local extremes, slope)
Global longitudinal analysis indices (activity, M, SD, recurrence)

Fig. 1. ReaderBench Processing Pipeline *(where M = Mean and SD = Standard Deviation)*

Course grade predictions were performed via a Recurrent Neural Network (RNN) model with LSTM cells considering the features from Fig. 1. Moreover, numerous sociograms were generated to depict interaction patterns between students and tutors, their evolution and behaviors. Results based on CNA indices and longitudinal analysis revealed that before COVID-19 (2018-2019), lower fluctuations in students' activity were noticed, in contrast to the pandemic period (2019-2020), which generated a strong increase in online participation with considerably more threads and more connected network, followed by a diminishing rate towards the end of the academic year. Also, the textual complexity indices denote a more elaborate and sophisticated discourse in the second academic year. The model trained for the second year during the pandemic explained more variance ($R^2 = .34$) than a model trained for the 2018-2019 academic year ($R^2 = .27$), which was expected since more information was readily available in Moodle.

Another research study [32] targeted the extraction and clustering of main ideas from student feedback. Feedback mechanisms for academic programs are frequently used to measure students' satisfaction, and open-text detailed impressions enable academics to improve their courses continually. Nonetheless, processing hundreds of student feedback messages across multiple subjects is time-consuming; hence, there is a need for an automated feedback summarizer capable of extracting the main ideas on various components for each educational program. The corpus used for the experiment consisted of 8,201 feedback responses for 168 distinct courses from the Computer Science Department of University Politehnica of Bucharest, corresponding to the 2019-2020 academic year. The method's workflow is presented in Fig. 2.

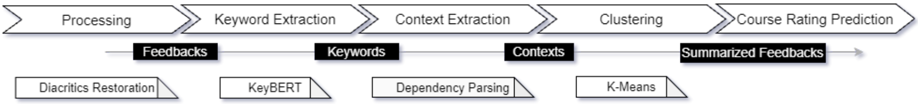

Fig. 2. Automated Feedback Processing Pipeline

Keywords were extracted from each course using RoBERT to find relevant contexts for frequently occurring concepts and group those contexts together using clustering techniques. During the preprocessing stage, three components were extracted from students' feedback: the course evaluation (a rating from 0 to 5), the positive aspects of the course, and potential improvement points. A diacritics restoration model [51] was applied to improve text quality, followed by a top 10 keywords extraction using KeyBERT [92]. The authors leveraged the Maximal Marginal Relevance method [93] to ensure the diversity of extracted keywords. The next step considered context extraction for each keyword and was performed in two steps. First, the sentence where the keyword appeared was captured; second, the dependency tree of each sentence was traversed only to extract relevant information related to the keyword. The last step leveraged k-Means clustering applied on keywords with more than 5 contexts to facilitate reading ease. The results argued that the method was efficient while reducing the overall text volume by 59%.

Additional various tasks were tackled while training Romanian open-source models with numerous applications in education, such as neural grammatical error corrections [51], diacritics restoration [94], text evaluations to improve student writing [5, 95], social media language analytics [57], ranking of news publications [96] or conversational agents [97]. Progress is also registered for speech-related tasks, namely automatic pronunciation assessment for Romanian [98] or improving multimodal speech recognition [99].

To sum up, the discussed educational NLP instruments dedicated to Romanian, built on state-of-the-art language models, introduced a transformative dimension to pedagogical practices. ReaderBench, for instance, is a comprehensive framework designed to enhance learning experiences through multilingual text analysis. Its robust modules, ranging from textual complexity evaluation to cohesion-based assessment, provide educators with invaluable insights into students' writing. Educators can leverage ReaderBench's capabilities for automated scoring and feedback. The extraction of the main ideas presented in the second study can empower teachers to improve their academic programs, identifying areas for clarification and thus enhancing course offerings. Hence, integrating these tools into the didactic process promotes data-driven decision-making and personalized education.

7 Conclusions

This paper highlights that using NLP tools has immense potential to enrich educational practices globally. Students can better understand language nuances and cultural contexts and communicate effectively by incorporating these tools into the school curricula. Embedding NLP-powered chatbots or virtual teaching assistants that engage students in real-time, answering queries or providing personalized guidance can alleviate teachers' load and promote continuous learning even outside the classroom. NLP-enhanced learning applications could empower students to engage with languages beyond their immediate environment, fostering a sense of global interconnectedness. Moreover, the advances in NLP are transforming the educational and academic environment by addressing various challenges and creating personalized learning experiences. The designed intelligent systems optimize didactic activities, ensuring personalized learning and feedback to match individual learning needs and contribute to student engagement and participation.

While most research in the domain is dedicated to high-resource languages like English, there is a need to develop tools for languages with limited coverage, such as Romanian. Through continuous investments in developing Romanian language resources, there has been significant progress in advancing dedicated language models and benchmarks, including corpora development. In this context, the current paper provides a comprehensive review of existing NLP Romanian instruments.

We showcase two research studies with applications in academic settings, namely assessing students' performance in Moodle courses and the extraction and clustering of main ideas from students' feedback, which aim to highlight the progress made in developing Romanian educational frameworks and their practical applications.

Future directions of development for Romanian NLP instruments may target the exploration of LLM applicability in educational contexts – for example, educational chatbots, automated question answering, automated grading, and language assessment. Furthermore, the development of additional resources, such as Natural Language Inference (NLI) for Romanian, should be considered to ensure a deeper understanding of the underlying texts.

The paper concludes by emphasizing the need for continued research and development in NLP instruments for low-resource languages such as Romanian, as the language is evolving beyond its low-resource status and is gaining more recognition in the global NLP community. The expansion in the research area shows promising results for the evolution of dedicated NLP instruments, going beyond a low-resource language. Nonetheless, this paper transcends its immediate linguistic focus to provide universal insights into the world of NLP. By presenting adaptable strategies, universal takeaways, and a vision for educational integration, this paper serves as inspiration for researchers, teachers, and practitioners seeking to leverage the transformative potential of NLP tools regardless of the targeted language.

This study opens pathways for future research that transcend language-specific boundaries. Researchers working on other low-resource languages can apply the methodologies outlined here to address challenges unique to their linguistic context. Furthermore, exploring techniques to enhance cross-lingual transfer learning or leveraging multilingual resources can yield insights with far-reaching implications.

References

1. Meurers, D.: Natural Language Processing and Language Learning. Encyclopedia of applied linguistics, 4193-4205 (2012) https://doi.org/10.1002/9781405198431.wbeal0858
2. Nadkarni, P., Ohno-Machado, L., Chapman, W.: Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18, 544-551 (2011) https://doi.org/10.1136/amiajnl-2011-000464
3. Yarlett, D.G., Ramscar, M.J.A.: Language Learning Through Similarity-Based Generalization. (2008)
4. Gu, P.Y.: Vocabulary learning in a second language: Person, task, context and strategies. TESL-EJ, 7(2), 1-25 (2003)

5. Florea, A.-M., Dascalu, M., Sirbu, M.-D., Trausan-Matu, S.: Improving Writing for Romanian Language. 4th Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2019), 131-141 (2019) https://doi.org/10.1007/978-981-13-9652-6_12
6. Lemaire, B., Mandin, S., Dessus, P., Denhière, G.: Computational cognitive models of summarization assessment skills. In: 27th Annual Conference of the Cognitive Science Society (CogSci' 2005). Erlbaum, Mahwah, NJ (2005)
7. Joshi, M., Rosé, C.P.: Using Transactivity in Conversation Summarization in Educational Dialog. In: SLaTE Workshop on Speech and Language Technology in Education, Farmington, Pennsylvania, USA (2007) https://doi.org/10.21437/SLaTE.2007-12
8. Barzilay, R., Elhadad, M.: Using lexical chains for text summarization. In: ACL Workshop on Intelligent Scalable Text Summarization (ISTS'97), pp. 10-17. ACL, Madrid, Spain (1997)
9. Duclaye, F., Yvon, F., Collin, O.: Learning paraphrases to improve a question-answering system. In: Proceedings of the EACL Workshop on Natural Language Processing for Question Answering Systems, pp. 35-41 (2003)
10. Oprescu, B., Dascalu, M., Trausan-Matu, S., Dessus, P., Bianco, M.: Automated Assessment of Paraphrases in Pupil's Self-Explanations. University Politehnica of Bucharest Scientific Bulletin Series C-Electrical Engineering and Computer Science, 76(1), 31-44 (2014)
11. Botarleanu, R., Dascalu, M., Sirbu, M.D., Crossley, S.A., Trausan-Matu, S.: Automated Text Simplification through Paraphrasing using Sequence-to-Sequence Models. In: 20th Int. Conf. on Artificial Intelligence in Education (AIED 2019). Springer, Chicago, IL (submitted)
12. Ruseti, S.: Advanced Natural Language Processing Techniques for Question Answering and Writing Evaluation. PhD Thesis, (2019)
13. Ravichandran, D., Hovy, E.: Learning surface text patterns for a question answering system. In: Proceedings of the 40th annual meeting on association for computational linguistics, pp. 41-47. Association for Computational Linguistics (2002) https://doi.org/10.3115/1073083.1073092
14. Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., Trausan-Matu, S.: Automated Essay Scoring in Applied Games: Reducing the Teacher Bandwidth Problem in Online Training. Computers & Education, 123, 212-224 (2018) https://doi.org/10.1016/j.compedu.2018.05.010
15. Dascalu, M., Westera, W., Ruseti, S., Trausan-Matu, S., Kurvers, H.: ReaderBench Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch. In: 18th Int. Conf. on Artificial Intelligence in Education (AIED 2017), pp. 52-63. Springer, Wuhan, China (2017) https://doi.org/10.1007/978-3-319-61425-0_5
16. McNamara, D.S., Crossley, S.A., Roscoe, R., Allen, L.K., Dai, J.: A hierarchical classification approach to automated essay scoring. Assessing Writing, 23, 35-59 (2015) https://doi.org/10.1016/j.asw.2014.09.002
17. Kinshuk: Developing adaptive and personalized learning environments. NY: Routledge, New York (2016) https://doi.org/10.4324/9781315795492
18. Botarleanu, R.-M., Dascalu, M., Sirbu, M.-D., Crossley, S.A., Trausan-Matu, S.: ReadME – Generating Personalized Feedback for Essay Writing using the ReaderBench Framework. In: 3rd Int. Conf. on Smart Learning Ecosystems and Regional Development (SLERD 2018), pp. 133-145. Springer, Aalborg, Denmark (2018) https://doi.org/10.1007/978-3-319-92022-1_12
19. Vidal-Abarca, E., Gilabert, R., Ferrer, A., Ávila, V., Martínez, T., Mañá, A., Llorens, A.C., Gil, L., Cerdán, R., Ramos, L., Serrano, M.A.: TuinLEC, an intelligent tutoring system to improve reading literacy skills / TuinLEC, un tutor inteligente para mejorar la competencia lectora. Infancia y Aprendizaje, 37, 25-56 (2014) https://doi.org/10.1080/02103702.2014.881657

20. Manning, C.D., Raghavan, P., Schütze, H.: Introduction to Information Retrieval, Vol. 1. Cambridge University Press, Cambridge, UK (2008) https://doi.org/10.1017/CBO9780511809071
21. Yuan, Z., Felice, M.: Constrained grammatical error correction using statistical machine translation. In: Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task, pp. 52-61 (2013)
22. Wagner, J., Foster, J., van Genabith, J.: A comparative evaluation of deep and shallow approaches to the automatic detection of common grammatical errors. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) (2007)
23. Trausan-Matu, S., Rebedea, T., Dascalu, M.: Analysis of discourse in collaborative learning chat conversations with multiple participants. In: Tufis, D., Forascu, C. (eds.) Multilinguality and Interoperability in Language Processing with Emphasis on Romanian, pp. 313-330. Editura Academiei Romane, Bucharest, Romania (2010)
24. Dessus, P., Trausan-Matu, S.: Implementing Bakhtin's dialogism theory with NLP techniques in distance learning environments. In: Trausan-Matu, S., Dessus, P. (eds.) Proc. 2nd Workshop on Natural Language Processing in Support of Learning: Metrics, Feedback and Connectivity (NLPsL 2010), pp. 11-20. Matrix Rom, Bucharest, Romania (2010)
25. Graesser, A.C., Rus, V., D'Mello, S., Jackson, G.T.: Autotutor: Learning through Natural Language Dialogue that Adapts to the Cognitive and Affective States of the Learner.
26. Graesser, A.C., Penumatsa, P., Ventura, M., Cai, Z., Hu, X.: Using LSA in AutoTutor: Learning through mixed-initiative dialogue in natural language. In: Landauer, T.K., McNamara, D.S., Dennis, S., Kintsch, W. (eds.) Handbook of Latent Semantic Analysis, pp. 243-262. Erlbaum, Mahwah, NJ (2007)
27. Crossley, S.A., McNamara, D.S.: Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. International Journal of Continuing Engineering Education and Life-Long Learning, 21(2/3), 170-191 (2011) https://doi.org/10.1504/IJCEELL.2011.040197
28. Peters, E., Hulstijn, J.H., Sercu, L., Lutjeharms, M.: Learning L2 German vocabulary through reading: The effect of three enhancement techniques compared. Language learning, 59(1), 113-151 (2009) https://doi.org/10.1111/j.1467-9922.2009.00502.x
29. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. Computer Speech and Language, 23, 89-106 (2009) https://doi.org/10.1016/j.csl.2008.04.003
30. Rebedea, T., Dascalu, M., Trausan-Matu, S.: PolyCAFe: Polyphony-based system for collaboration analysis and feedback generation. In: Second Workshop on Natural Language in Support of Learning: Metrics, Feedback and Connectivity, pp. 21-34. MatrixRom, Bucharest, Romania (2010)
31. Niculescu, M.A., Ruseti, S., Dascalu, M.: RoSummary: Control Tokens for Romanian News Summarization. Algorithms 2022 (MDPI), 15(472), N/A (2022) https://doi.org/10.3390/a15120472
32. Masala, M., Ruseti, S., Dascalu, M., Dobre, C.: Extracting and Clustering Main Ideas from Student Feedback using Language Models. Proceedings of 22nd Int. Conf. on Artificial Intelligence in Education (AIED 2021), 12748, 282-292 (2021) https://doi.org/10.1007/978-3-030-78292-4_23
33. Toma, I., Marica, A.M., Dascalu, M., Trausan-Matu, S.: Readerbench – Automated Feedback Generation for Essays in Romanian. U.P.B. Sci. Bull. Series C – Electrical Engineering and Computer Science, 83(2), 21-34 (2021)
34. Deane, P.: On the relation between automated essay scoring and modern views of the writing construct. Assessing Writing, 18, 7-24 (2013) https://doi.org/10.1016/j.asw.2012.10.002
35. Foltz, P.W., Laham, D., Landauer, T.K.: Automated essay scoring: applications to Educational Technology. Int. Conf. ED-MEDIA '99, Seattle (1999)

36. Panaite, M., Dascalu, M., Johnson, A.M., Balyan, R., Dai, J., McNamara, D.S., Trausan-Matu, S.: Bring it on! Challenges Encountered while Building a Comprehensive Tutoring System using ReaderBench. In: 19th Int. Conf. on Artificial Intelligence in Education (AIED 2018). Springer, London, UK (2018) https://doi.org/10.1007/978-3-319-93843-1_30
37. Jugo, I., Kovačić, B., Tijan, E.: Cluster analysis of student activity in a web-based intelligent tutoring system. Scientific Journal of Maritime Research, 29, 75-83 (2015)
38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I.: Attention Is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Long Beach, California, USA (2017) 6000-6010
39. Trandabat, D., Irimia, E., Mititielu, V.B., Cristea, D., Tufis, D.: The Romanian Language in the Digital Era. Springer, Metanet White Paper Series, (2012)
40. Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J.: The MARCELL Legislative Corpus. Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France (2020) 3761-3768
41. Pais, V., Mitrofan, M., Gasan, C.L., Ianov, A., Ghita, C., Coneschi, V.S., Onut, A.: Romanian Named Entity Recognition in the Legal domain (LegalNERo). Zenodo, (2021) https://doi.org/10.18653/v1/2021.nllp-1.2
42. Rebeja, P., Chitez, M., Rogobete, R., Dinca, A., Bercuci, L.: ParlaMint-RO: Chamber of the Eternal Future. Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France (2022) 131-134
43. Mitrofan, M., Tufis, D.: BioRo: The Biomedical Corpus for the Romanian Language. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018)
44. Mitrofan, M., Barbu Mititelu, V., Mitrofan, G.: MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language. Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, Florence, Italy (2019) 71-79 https://doi.org/10.18653/v1/W19-5008
45. Mitrofan, M., Pais, V.: Improving Romanian BioNER Using a Biologically Inspired System. Proceedings of the 21st Workshop on Biomedical Language Processing. Association for Computational Linguistics, Dublin, Ireland (2022) 316-322 https://doi.org/10.18653/v1/2022.bionlp-1.30
46. Mititelu, V.B., Mitrofan, M.: The Romanian Medical Treebank – SiMoNERo. Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR 2020) (2020) 7-16
47. Tufis, D., Barbu, E., Mititelu, V.B., Ion, R., Bozianu, L.: The Romanian Wordnet. Romanian Journal of Information Science and Technology (ROMJIST), 7, 107-124 (2004)
48. Ion, R., Irimia, E., Stefanescu, D., Tufis, D.: ROMBAC: The Romanian Balanced Annotated Corpus. Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), (2012)
49. Dumitrescu, S.D., Avram, A.M.: Introducing RONEC – the Romanian Named Entity Corpus. Proceedings of the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France (2020) 4436-4443
50. Aleris: RoTex Corpus Builder. https://github.com/aleris/ReadME-RoTex-Corpus-Builder, last accessed 07/17/2023
51. Cotet, T.-M., Ruseti, S., Dascalu, M.: Neural grammatical error correction for romanian. IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), Baltimore, MD, USA (2020) 625-631 https://doi.org/10.1109/ICTAI50040.2020.00101

52. Mititelu, V.B., Irimia, E., Tufis, D.: The Reference Corpus of Contemporary Romanian Language (CoRoLa). Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018) 1235-1239
53. Butnaru, A.M., Ionescu, R.T.: MOROCO: The Moldavian and Romanian Dialectal Corpus. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, Florence, Italy (2019) 688-698 https://doi.org/10.18653/v1/P19-1068
54. Hoefels, D.C., Çöltekin, Ç., Mădroane, I.D.: CoRoSeOf – An Annotated Corpus of Romanian Sexist and Offensive Tweets. Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC'22). European Language Resources Association, Marseille, France (2022) 2269-2281
55. Maonlescu, M., Çøltekin, Ç.: Roff – A Romanian Twitter Dataset for Offensive Language. Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). INCOMA Ltd., Online (2021) 895-900 https://doi.org/10.26615/978-954-452-072-4_102
56. Paraschiv, A., Sandu, I., Cercel, D.-C., Dascalu, M.: Fighting Romanian Offensive Language with RO-Offense: A Dataset and Classification Models for Online Comments. Preprint submitted to Elsevier, (2022)
57. Cojocaru, A., Paraschiv, A., Dascalu, M.: News-RO-Offense – A Romanian Offensive Language Dataset and Baseline Models Centered on News Article. Proceedings of RoCHI 2022, (2022) https://doi.org/10.37789/rochi.2022.1.1.12
58. Tufis, D., Irimia, E.: RoCo-News: A Hand Validated Journalistic Corpus of Romanian. Proceedings of the Fifth International Conference on Language Resources and Evaluation (2006) 869-872
59. Artetxe, M., Ruder, S., Yogatama, D.: On the Cross-lingual Transferability of Monolingual Representations. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online (2020) 4623-4637 https://doi.org/10.18653/v1/2020.acl-main.421
60. Dumitrescu, S.D., Rebeja, P., Lorincz, B., Gaman, M., Avram, A., Ilie, M., Pruteanu, A., Stan, A., Rosia, L., Iacobescu, C., Morogan, L., Dima, G., Marchidan, G., Rebedea, T., Chitez, M., Yogatama, D., Ruder, S., Ionescu, R.T., Pascanu, R., Patraucean, V.: LiRo: Benchmark and leaderboard for Romanian language tasks. In: Vanschoren, J., Yeung, S. (eds.): Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks, Vol. 1. Curran (2021) N/A
61. Nicolae, D.C., Tufis, D.: RoITD: Romanian IT Question Answering Dataset. ConsILR-2021, 1154-1161 (2019)
62. Avram, S.M., Oltean, M.: A comparison of several AI techniques for authorship attribution on Romanian texts. arXiv:2211.05180 Mathematics 2022, 4589 (2022)
https://doi.org/10.3390/math10234589
63. Oravițan, A., Chitez, M., Bercuci, L., Rogobete, R.: Using the bilingual Corpus of Romanian Academic Genres (ROGER) platform to improve students' academic writing. Intelligent CALL, granular systems and learner data: short papers from EUROCALL 2022, 315-321 (2022) https://doi.org/10.14705/rpnet.2022.61.1477
64. Javier Ortiz Suarez, P., Sagot, B., Romary, L. and Sagot, B.B.: Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), (2019)
65. Cristea, D., Pistol, I., Boghiu, S., Bibiri, A.D., Gifu, D., Scutelnicu, A., Onofrei, M., Trandabat, D., Bugeag, G.: CoBiLiRo: A Research Platform for Bimodal Corpora. Proceedings of the 1st International Workshop on Language Technology Platforms, Marseille, France, 22-27 (2020)
66. Delvin, J., Chang, M.-W,, Lee, K. and Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1, 4171-4186 (2019)

67. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving Language Understanding by Generative Pre-Training. Technical Report, OpenAI, (2018)
68. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. Technical Report, OpenAI, (2019)
69. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V.: XLNet: Generalized Autoregressive Pretraining for Language Understanding. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS 2019, 5754-5764 (2019)
70. Masala, M., Ruseti, S. and Dascalu, M.: RoBERT-A Romanian BERT Model. COLING, 6626-6637 (2020) https://doi.org/10.18653/v1/2020.coling-main.581
71. Dumitrescu, S.D., Avram, A.M., Pyysalo, S.: The birth of Romanian BERT. Findings of the Association for Computational Linguistics: EMNLP 2020, 4324-4328 (2020) https://doi.org/10.18653/v1/2020.findings-emnlp.387
72. Tiedemann, J.: Parallel data, tools and interfaces in opus. LREC, (2012)
73. Niculescu, M.A., Ruseti, S., Dascalu, M.: RoGPT2: Romanian GPT2 for Text Generation. 33rd International Conference on Tools with Artificial Intelligence (ICTAI), 1154-1161 (2021) https://doi.org/10.1109/ICTAI52525.2021.00183
74. Radford, A., Wu, J., Rewon, C., Luan, D., Amodei, D., Sutskever, I.: Language Models are Unsupervised Multitask Learners. OpenAI Blog, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf, Online (2019)
75. Buzea, M., Trausan-Matu, S., Rebedea, T.: Automatic Romanian Text generation using GPT-2. U.P.B. Sci. Bull. Series C, 84(4), 15-30 (2022)
76. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT. ICLR2020, arXiv:1904.09675, (2020)
77. Papieni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA, Philadelphia, Pennsylvania, USA (2002) 311-318 https://doi.org/10.3115/1073083.1073135
78. Lin, C.: Recall-oriented understudy for gisting evaluation (rouge). (2005)
79. Dumitrscu, S., Mihai, I.: GPT-Neo Romanian 780M. GitHub Repository of Romanian-Transformers: https://github.com/dumitrescustefan/Romanian-Transformers (2022)
80. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual Denoising Pre-training for Neural Machine Translation. Transactions of the Association for Computational Linguistics, 8, 726-742 (2020) https://doi.org/10.1162/tacl_a_00343
81. Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Yepes, A.J., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., Zampieri, M.: Findings of the 2016 Conference on Machine Translation (WMT16). Proceedings of the First Conference on Machine Translation, 2, 131-198 (2016) https://doi.org/10.18653/v1/W16-2301
82. Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C.: mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 483-498 (2021) https://doi.org/10.18653/v1/2021.naacl-main.41
83. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Shane Gu, S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J.,

Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling Instruction-Finetuned Language Models. arXiv:2210.11416 [cs.LG], (2022)

84. Ranathunga, S., A de Silva, N.: Some Languages are More Equal than Others: Probing Deeper into the Linguistic Disparity in the NLP World. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, Vol. Volume 1: Long Papers, pp. 823-848. Association for Computational Linguistics (2022)
85. Khan, M.Z.: Comparing the Performance of NLP Toolkits and Evaluation measures in Legal Tech. ArXiv, abs/2103.11792 (2021)
86. Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S., Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., Søgaard, A.: Challenges and Strategies in Cross-Cultural NLP. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 6997-7013. Association for Computational Linguistics, Dublin, Ireland (2022) https://doi.org/10.18653/v1/2022.acl-long.482
87. Aji, A.F., Winata, G.I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R.E., Baldwin, T., Lau, J.H., Ruder, S.: One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Vol. Volume 1: Long Papers, pp. 7226-7249. Association for Computational Linguistics, Dublin, Ireland (2022) https://doi.org/10.18653/v1/2022.acl-long.500
88. Păiș, V., Ion, R., Tufiș, D.: A Processing Platform Relating Data and Tools for Romanian Language. Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP), 81-88 (2020)
89. Dascalu, M., Dessus, P., Trausan-Matu, S., Bianco, M., Nardy, A.: ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. International Conference on Artificial Intelligence in Education (AIED), 379-388 (2013) https://doi.org/10.1007/978-3-642-39112-5_39
90. Dascalu, M.D., Ruseti, S., Dascalu, M., McNamara, D.S., Carabas, M., Rebedea, T., Trausan-Matu, S.: Before and during COVID-19: A Cohesion Network Analysis of students' online participation in moodle courses. Computers in Human Behavior, 121, 106780-106780 (2021) https://doi.org/10.1016/j.chb.2021.106780
91. Dascalu, M., McNamara, D.S., Trausan-Matu, S., Allen, L.K.: Cohesion Network Analysis of CSCL Participation. Behavior Research Methods, 50(2), 604-619 (2018) https://doi.org/10.3758/s13428-017-0888-4
92. Grootendorst, M.: KeyBERT: minimal keyword extraction with BERT. https://github.com/MaartenGr/KeyBERT, (2020)
93. Carbonell, J., Goldstein, J.: Use of MMR, diversity-based reranking for reordering documents and producing summaries. SIGIR Forum (ACM Special Interest Group on Information Retrieval), 335-336 (1998) https://doi.org/10.1145/290941.291025
94. Ruseti, S., Cotet, T.-M., Dascalu, M.: Romanian Diactrics Restoration using Recurrent Neural Networks. ArXiv, abs/2009.02743 (2020)
95. Sirbu, M.-D., Dascalu, M., Gifu, D., Cotet, T.-M., Tosca, A., Trausan-Matu, S.: ReadME – Improving Writing Skills in Romanian Language. In: Pais, V., Gifu, D., Trandabat, D., Cristea, D., Tufis, D. (eds.): Proceedings of the 13th Int. Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018), Iasi, Romania (2018) 135-145
96. Busuioc, C., Ruseti, S., Dascalu, M.: A Literature Review of NLP Approaches to Fake News Detection and Their Applicability to Romanian-Language News Analysis. Tansilvania, 65-71 (2018) https://doi.org/10.51391/trva.2020.10.07

97. Boroghina, G., Corlatescu, D.-G., Dascalu, M.: Conversational Agent in Romanian for Storing User Information in a Knowledge Graph. International Conference on Human-Computer Interaction (RoCHI2020), (2020) https://doi.org/10.37789/rochi.2020.1.1.15
98. Ungureanu, D., Ruseti, S., Toma, I., Dascalu, M.: pRonounce: Automatic Pronounciation Assessment for Romanian. Conference on Smart Learning Ecosystems and Regional Development (SLERD 2022), Polyphonic Construction of Smart Learning Ecosystems, 103-114 (2022) https://doi.org/10.1007/978-981-19-5240-1_7
99. Oneata, D., Cucu, H.: Improving Multimodal Speech Recognition by Data Augmentation and Speech Representations. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) New Orleans, LA, USA (2022) 4578-4587 https://doi.org/10.1109/CVPRW56347.2022.00504