ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **24**, Number 4, 2021, 384–401

In-depth evaluation of Romanian natural language processing pipelines

Vasile Păiș , Radu Ion , Andrei-Marius Avram , Maria Mitrofan , and Dan Tufiș

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
Email: vasile,radu,andrei.avram,maria,tufis@racai.ro

Abstract. With the increased size of Universal Dependencies tree banks, several basic language processing kits (BLARK) for multiple languages appeared in recent years, indicating improved performances on different languages. Nevertheless, published results are not directly comparable for the Romanian language since different tools make use of different Universal Dependencies versions and different additional resources, such as pre-trained word embeddings. In this paper, we re-train several state-of-the-art tools for processing Romanian language by using a common methodology comprising of training and evaluating on the same version of RoRefTrees corpus and using the same pre-trained word embeddings from the representative corpus of contemporary Romanian language (CoRoLa). Furthermore, we also explore the capabilities of the trained models when faced with unseen text from a different domain. For this purpose, we further test the resulting model on the SiMoNERo corpus. We employ different metrics to assess the performance on operations like tokenization, sentence splitting, lemmatization, part-of-speech tagging and dependency parsing.

Keywords: Natural language processing; BLARK; performance evaluation;Romanian text processing

1 Introduction

Natural language understanding, from the artificial intelligence point of view, entails, as a first step, the transformation of raw text into a structured and annotated format with elements such as sentence boundaries, tokens, lemmata, part-of-speech tags, dependency parsing, named entities and any other additional features useful for higher level algorithms. This process is usually handled by multiple modules combined into an annotation pipeline, where each module makes use of the output produced by previous modules within the pipeline. The development of specialized NLP modules and chaining them into pipelines was a constant preoccupation at the Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy since its inception. After a long period of dominance of the knowledge and unification (rule) based approaches (which resulted in advanced processing platforms such as IURES [26], MACPAIL [29] or EGLU [28,31] - a generalized multi-platform version of the then state-of-the-art unification environment ELU [27]), the data-driven statistical approaches became prevalent and our work followed that trend. As such, corpus linguistics research and language resource development took

over our priorities. The first NLP tools following the modern architectures were implemented within the Romanian Academy research programs (LINGUASTAT), and further enhanced during the TELRI project[1]. The QTAG was one of the first statistical taggers openly available and the first one for Romanian language [32–35]. The next generation tagger was TTL [8], an enhanced re-implementation of TnT [6], and which, for more than 10 years, was (and still is) the standard processing chain for tokenization, tagging and lemmatization in Romanian. With the advent of neural networks based approaches, new implementation of NLP & speech processing platforms were developed [36], the most recent one being RELATE [15, 37, 38].

A constant effort of our research group was directed towards the development of reference language resources, with wide coverage and as high accuracy as possible. Most of the basic resources for Romanian language have been manually validated (and some of them were even entirely manually created) and diligently maintained over years. Romanian WordNet is probably the best-known language resource we developed and was largely publicised [39–54]. The first electronic corpora for Romanian language, conforming with the international standards and best practices, were created beginning with the end of 90's and continues to this day. The corpus "1984" [55] created during the MULTEXT-East project [56], was the major gold training data for the first taggers for Romanian, as it was annotated by hand. Later, it was followed by ROMBAC [57] the first Romanian balanced corpus, publicly released and several other specialized corpora (ROCO-News [58], RoTime-Bank [59], Legal Corpus [23], Dependency Treebank [60, 61], Biomedical corpus MoNERo [62], Literary Corpus [63], Named entity corpus LegalNERo [85], SiMoNERo [3], RoRefTrees treebank (RRT) [87], etc.). For multilingual studies (especially in translation projects) we annotated several corpora, internationally acknowledged: JRC-Acquis [64–66], ACCURAT corpus [67], MARCELL corpus [68, 69].

The most important corpus, developed at our institute is CoRoLa (The Reference Corpus of **CO**ntemporary **RO**manian **LA**nguage), both in terms of size (more than 1 billion tokens) and content and quality. It is a bi-modal corpus (text and speech) with several friendly interfaces allowing a user to search for desired information both in the text and speech content. It and its successive versions have been largely presented in numerous papers at international conferences and in reviewed journals [12, 70–74]. The CoRoLa corpus is part of a pan-european initiative (EuReCo) to develop a multilingual reference corpus made of several national reference corpora [75]. Besides text corpora, we developed several speech data corpora (containing aligned speech records and their transcriptions) [76–79]. Some of these speech corpora have been included into the oral part of CoRoLa corpus, others are included in the CoBiLiRo speech data repository [80].

We only referred to text processing tools, but the results in speech processing are also important: the SpeeD group at University "Politehnica" of Bucharest[2] is a leader of Romanian research and development in the area of speech recognition (and not only). The "Speech Processing Group" at the Technical University of Cluj-Napoca[3] is another leader, mainly in expressive text-to-speech synthesis and speech corpora creation. Significant research activities in the areas of signal processing with fuzzy systems and neural networks, with application in medicine, speech analysis are also carried at the Institute for Computer Science in Iași[4]. The NLP group of University "A.I. Cuza" of Iași is also a relevant actor in the domain of language technology[5], specialized lately in the storage, management and exploitation of bimodal resources. They are the

[1] http://telri.nytud.hu/
[2] https://speed.pub.ro/
[3] https://speech.utcluj.ro/
[4] http://iit.academiaromana-is.ro/
[5] http://nlptools.info.uaic.ro/

authors of the CoBiLiRo platform [80], a sophisticated repository of the speech and text corpora produced by SpeeD group at University "Politehnica" of Bucharest, "Speech Processing Group" at the Technical University of Cluj-Napoca and the Research Institute for Artificial Intelligence "M. Drăgănescu" of the Romanian Academy within the ReTeRom project[6].

As the Linguistically Linked Open Data (LLOD) got impetus, our most significant language resources are converted to LLOD format [81] and some of them are included in the Linked Open Data Cloud[7]. Additionally, certain new resources, such as the Romanian Named Entity Recognition in the Legal domain (LegalNERo) corpus [85] are directly developed with associated RDF format files, specific to LLOD initiatives. Furthermore, SPARQL query interfaces are exposed in the RELATE platform[8] to directly interact with Romanian LLOD resources[9].

Transformer-based [83] pretrained language models have become ubiquitous in natural language processing, being applied to a wide variety of tasks. One of the most popular architecture is entitled Bidirectional Encoders from Transformer Representations (BERT) [86] and it uses multiple Transformer encoders that are stacked on top of each other. The model was trained on an English corpus composed of 3.3 billion words by using two self-supervised tasks: (1) Masked Language Modeling (MLM) - predict a percentage of the input tokens that were masked and (2) Next Sentence Prediction (NSP) - predict whether from two sentences `A` and `B` (in this order), `B` is the actual sentence that follows `A`. Moreover, a multilingual version of BERT (mBERT) [10] was released which was trained on more than 100 languages, including Romanian.

To the best of our knowledge, there exist two versions of BERT trained on Romanian corpora that were introduced in [82] and [84]. The first version - BERT-base-ro - was trained on 15.1 GB of Romanian text and was evaluated on three tasks: (1) simple UD - one model for UPOS and XPOS, (2) joint UD - one model for trained jontly on all tasks in UD and (3) named entity recognition. The latter version - RoBERT - comes in three variants (i.e. RoBERT-small, RoBERT-base and RoBERT-large) that were trained on 12.6 GB of Romanian text. The performance was evaluated also on three tasks, all different from the previous ones: (1) sentiment analysis, (2), Moldavian vs. Romanian Dialect and cross-dialect topic identification and (3) diacritics restoration. Both models reported an improvement in performance over multilingual language models such as mBERT. Even though multiple BERT models for Romanian language exist, there is no actual BLARK system developed using these models. Therefore, in this paper we are able to present only preliminary results using BERT models.

The goal of the European Language Equality (ELE) project[11] is the development of a sustainable evidence-based strategic research agenda and road map setting out actions, processes, tools, and actors to achieve full digital language equality of all languages (official or otherwise) used within the European Union through the effective use of language technologies. Furthermore, the research agenda encompasses the determination of the current state of language technologies and language equality within the EU. In this context, the present work focuses on evaluating existing Romanian basic language resource kits (BLARK) processing pipelines from multiple perspectives: performance of tokenization, sentence segmentation, lemmatization, and part-of-speech tagging; time required for training and annotation; additional features available from using a specific pipeline.

[6] https://www.racai.ro/p/reterom/index_en.html
[7] https://lod-cloud.net/
[8] https://relate.racai.ro/datasets/dataset.html
[9] https://www.racai.ro/p/llod/
[10] https://github.com/google-research/bert/blob/master/multilingual.md
[11] http://www.european-language-equality.eu

Lately, Romanian became a language addressed by several pipelines, some of them developed outside Romania and it was quite challenging to perform a honest and professional in-depth evaluation of the most relevant pipelines. However, we considered only projects where the training mechanism was freely available, to allow us to train all the systems on the same data sets (for this reason we did not consider commercial applications). Finally, we considered only systems accompanied by a research paper, describing the method used.

This paper is organized as follows: Section 2 describes the methodology employed throughout this work, Section 3 presents the annotation pipelines under evaluation, Section 4 presents the results and finally the conclusions are given in Section 5.

2 Methodology

The purpose of evaluation is to provide an assessment of the value of a solution to a given problem [13]. In this case, the purpose is to determine the performance of the different annotation levels provided by Romanian language processing pipelines. However, not all annotations are the same. Sentence splitting and tokenization aim at identifying boundaries in text for each category. Lemmatization will transform an already identified token into its corresponding lemma. Part of speech tagging will associate tags to tokens. Therefore, for each case different performance measures must be taken into consideration. Evaluation of boundary detection for tokens and sentences is handled similar to Jurish and Würzner (2013) [10]. Considering a set of boundaries *Brelevant* from a gold annotation corpus and a set of automatically detected boundaries *Bretrieved* from an application output, we define:

$$P = \frac{|B_{relevant} \bigcap B_{retrieved}|}{|B_{retrieved}|}$$

$$R = \frac{|B_{relevant} \bigcap B_{retrieved}|}{|B_{relevant}|}$$

In this case $|B_{relevant} \bigcap B_{retrieved}|$ denotes the number of correct boundaries predicted by the application. Therefore, the precision is defined as the ratio of correct boundaries from the total predicted boundaries, while the recall is defined as the ratio of correct boundaries from the total gold annotations. Furthermore, the F1 measure can be defined in the usual way:

$$F1 = \frac{2PR}{P + R}$$

Regarding part-of-speech tagging, accuracy seems to be the most intuitive and the most used performance measure mentioned in the literature [13]. Considering the set of correctly identified tags *Tcorrect* (a tag is considered correct if for a given token the predicted value matches with the gold annotation) and the set of all predicted tags T, the accuracy can be defined as:

$$Acc = \frac{|T_{correct}|}{|T|}$$

Similarly, for lemmatization we use the same accuracy measure, considering a lemma to be correct if the predicted value matches the gold standard annotation. It must be noted the observation of Adda et al. (1999) [1] concerning the word segmentation convention used by the tagger which must be the same as the one used for the reference data, otherwise there is a need

to deploy realignment procedures. To avoid realigning data, which could introduce errors, we evaluate separately the tokenization models and the other annotation models. For part-of-speech tagging and lemmatization we use the gold tokens as input to the annotator, thus eliminating the alignment issue.

Furthermore, Adda et al. (1999) [1] observes that the tagset used by the tagger must be the same as the one used to annotate the reference data, otherwise specific mapping procedures need to be applied. This is however not an issue in our case since each model is trained on training data which uses the same tagset as the testing data.

For tools supporting dependency parsing we considered the unlabeled attachment score (UAS), which checks for the correct head, and labeled attachment score (LAS), which provides the percentage of tokens for which the system has predicted both the correct head and the correct dependency relation.

Training is performed using the Universal Dependencies (UD) [12] version 2.7 of RoRefTrees treebank (RRT) [13]. It contains 9,523 sentences, manually validated at the syntactic level, that reflects the contemporary language, covering a variety of genres. This dataset contains both raw text and annotated gold data in CoNLL-U format. The annotated data is already split into three subsets: training, dev(elopment) and test. We first explore the different models' performance on the same dataset, by using the splits of the RRT dataset. Afterwards, we train a new model on the complete RRT dataset (by combining all the splits) and explore how well the models adapt when exposed to a completely new dataset, from a different domain. For this purpose, we use SiMoNERo [3] which is a Romanian bio-medical treebank [14] that is also annotated with Universal Dependencies syntactic relations and is available within UD version 2.7. SiMoNERo is the first Romanian medical treebank. It was buit on a gold standard morphologically annotated corpus, also containing manually validated annotation with medical named entities. SiMoNERo treebank contains 4,681 sentences split into three files: training, development and test. A characteristic of this treebank is that the average sentence length, 31.19 tokens/sentence, is above 22.94 tokens/sentence, the average sentence length in RRT treebank. Besides their different sizes, when comparing the vocabulary specific to each treebank, 58% of the unique lemmas in SiMoNERo do not occur in RRT. They are mainly medical terms such as: infarct (heart attack), neuroglicopenic (neuroglycopenic), osteoblastic (osteoblastic), etc.

Table 1: General statistics between RRT and SiMoNERo.

	RRT	**SiMoNERo**
No. sentences	9,523	4,681
Tokens	218,511	146,020
Unique lemmas	17,458	10,711
Tokens/Sentence	22.94	31.19
Punctuation/Sentence	2.9	4.2

Since the corpora used contains both Universal Part-of-Speech[15] (UPOS) tags and language specific tags (XPOS) we evaluate the model performance on both tag categories. There is a direct mapping from XPOS to UPOS and it was previously found [22] that for some statistical

[12] https://universaldependencies.org/
[13] https://github.com/UniversalDependencies/UD_Romanian-RRT
[14] https://universaldependencies.org/treebanks/ro_simonero/index.html
[15] https://universaldependencies.org/u/pos/

algorithms a better performance is achieved by training on a larger tagset. Therefore, when possible, we also explore the variant of training on XPOS and then obtaining the UPOS tags via mapping to check if this observation still holds for current state-of-the art neural algorithms.

To give a fair comparison in terms of time required for training and running for the different systems, we used for our experiments the same server having two Intel Xeon Silver 4210 CPUs (each with 10 physical cores and 20 threads) running at 2.2 Ghz and a total of 128 Gb RAM. Even though certain tools under evaluation could benefit from an increased number of CPUs or even GPU acceleration, we limited our evaluation to this single environment. Additionally, even though some systems can benefit from fine-tuning certain parameters, we only used the default system settings. This choice was justified because parameter fine-tuning usually accounts for marginal improvements at the expense of making the model more dependent on receiving similar testing data, while for our experiments we also wanted to investigate the application of the generated models on out-of-domain data (the SiMoNERo dataset).

For systems making use of pre-trained word embeddings, we used the embeddings described in [14] trained on the Representative Corpus of Contemporary Romanian Language (CoRoLa) [12, 23], following the approach of Bojanowsky et al. (2017) [4]. This allowed us to evaluate the algorithms in similar conditions.

With the introduction of Universal Dependencies (UD) data for training language models in many languages, the CoNLL-U format becomes more popular for providing annotations in a standardized format. Therefore, we also explore in this work the suitability of the studied tools for providing CoNLL-U formatted output files.

3 Annotation pipelines

In this section we present the key characteristics of the different basic language processing kits that were analysed for the purposes of this work. Their presentation is given following the year of publication for the paper describing each system.

TreeTagger[16] [17] is a tool for annotating text with part-of-speech and lemma information. Developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart, the TreeTagger uses decision trees to estimate transition probabilities for part of speech tagging. Tokenization is handled outside the model, by means of a rule-based approach, thus requiring no training. However, sentence segmentation is handled within the model, by using a special tag "SENT" assigned to tokens indicating the end of sentence. Apart from the annotated training data, it makes use of a lexicon containing the possible tags and lemmas identifiable in the training data associated with each word form. We use the training data to generate this information. Once this is provided together with the training data, the TreeTagger can be trained to produce an annotation model. A single type of tag can be predicted by a model, thus requiring two different models: one for UPOS and one for XPOS. Training and results are using a column format. However, this is not CoNLL-U, therefore converters were implemented to transform between the specific format and CoNLL-U. Finally, a major shortcoming of the TreeTagger is the fact that it can only produce lemmas for words present in the training set, by using the previously created lexicon. There is no algorithm for creating lemmas for unseen words. Thus, it is estimated that performance for lemmatization can be improved by employing a larger lexicon, built using additional resources. Nevertheless, for the purposes of the current

[16] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

work, we limited the lexicon creation to the available training data in order to make the results comparable to those offered by other tools.

TTL [8] is a Perl 5 module developed at RACAI that performs sentence splitting, tokenization, POS tagging, lemmatization and chunking for four languages: Romanian, English, French and Bulgarian. Sentence splitting and tokenization are driven by rule-based algorithms, using lists of multi-word expressions (e.g. "de la", "cel puțin"), abbreviations (e.g. "etc.", "dl."), and clitic splits (e.g. "-ți", "-mi" or "mi-") to recognize complex tokens and to split clitics off words that have them attached and which have a morpho-syntactic interpretation of their own. POS tagging is performed via HMM modeling, based on Brants' TnT [6], improving on unknown words heuristics for Romanian. Lemmatization uses a Markov Model to estimate the probability of a lemma candidate, automatically inferred by applying string transformation rules, using the longest common substring, computed between word forms and their respective lemmas in the lexicon. Finally, chunking (the detection of non-recursive noun, verb, adjectival and adverbial phrases) is realized with regular expressions over sequences of POS tags. A version of this tool is available from the RETEROM[17] project's GitHub repository[18].

UDPipe[19] [20] is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. Developed at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic, it can be trained given annotated data in CoNLL-U format. This removes the need for any converters between the UD data and UDPipe input/output. UDPipe uses an artificial neural network for tokenization. This is followed by POS tagging and lemmatization, using MorphoDiTa [21]. The POS tagger is based in part on the work of Spoustova et al. (2009) [19] and is implemented as a supervised, feature-rich averaged perceptron [7], employing dynamic programming at runtime (Viterbi decoder). Finally, UDPipe employs a dependency parser described in [25], capable of parsing both projective and non-projective sentences using a neural network classifier for prediction, requiring no feature engineering.

NLP-Cube[20] [5] is a framework allowing end-to-end text processing with neural networks, developed by researchers at Adobe and RACAI. It performs sentence splitting, tokenization, compound word expansion, lemmatization, tagging and parsing. For each function, a dedicated model must be trained. Like UDPipe, it makes use of data in CoNLL-U format, thus removing the need for specific converters. It is written in Python and is based entirely on recurrent neural networks built in DyNET [11]. It makes use of pre-trained word embeddings, therefore, as mentioned in the Methodology section, we employed the embeddings [14] trained on the CoRoLa corpus.

RNNTagger[21] [18] was implemented in Python using the Deep Learning library PyTorch. It employs a character-based BiLSTM tagger for part-of-speech tagging and a character-based encoder-decoder architecture for lemmatization. Like the TreeTagger, it needs separated models for UPOS and XPOS tags, but due to the neural algorithm employed, lemmatization is possible regardless if the word was seen or not during training. Tokenization is handled similarly to TreeTagger using a rule-based approach, though different from the TreeTagger tokenizer. It employs pre-trained word embeddings but requires input/output converters for the CoNLL-U format.

[17] https://www.racai.ro/p/reterom/
[18] https://github.com/racai-ai/TEPROLIN
[19] https://github.com/ufal/udpipe
[20] https://github.com/adobe/NLP-Cube
[21] https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/

Stanza[22] [16] is a Python natural language analysis package. As indicated by the authors, it is built with neural network components that enable efficient training and evaluation using annotated data. In addition, Stanza includes a Python interface to the Stanford CoreNLP Java package and inherits additional functionality from there, such as constituency parsing, coreference resolution, and linguistic pattern matching. It employs the CoNLL-U format at both training and runtime and makes use of pre-trained word embeddings.

4 Results

As described in the Methodology section, we trained and evaluated each tool twice. First, on the RRT corpus (considering the train, dev, and test splits accordingly) and second, each tool was trained on the entire RRT corpus (all splits combined into a single train corpus) and evaluated on the SiMoNERo corpus. Results from the evaluation on the RRT corpus are presented in Table 2, while results from evaluation on SiMoNERo are given in Table 3. We also present some preliminary results on UPOS and XPOS tagging with BERT-base-ro and RoBERT variants for the first evaluation in Table 4 for the RRT corpus, using the scripts provided by the authors of BERT-base-ro[23].

Table 2: Performance comparison of different models trained and evaluated against the RRT corpus for tokenization, sentence splitting, lemmatization, part-of-speech tagging.

	Tokens F1	**Sentences F1**	**Lemma Acc**	**UPOS Acc**	**XPOS Acc**
TreeTagger	99.72	98.37	89.62	95.58	92.97
TTL	99.66	96.19	95.84	95.54	95.18
UDPipe	99.88	97.39	95.91	97.15	96.24
NLP-Cube	99.86	**98.62**	51.39	97.73	96.94
RNNTagger	99.23	95.47	98.30	**97.82**	**97.19**
Stanza	**99.94**	98.42	**98.90**	97.64	97.07

Table 3: Performance comparison of different models trained on RRT and evaluated against the SiMoNERo corpus for tokenization, sentence splitting, lemmatization, part-of-speech tagging.

	Tokens F1	**Sentences F1**	**Lemma Acc**	**UPOS Acc**	**XPOS Acc**
TreeTagger	99.86	98.11	81.89	93.64	88.05
TTL	99.85	96.24	95.67	92.15	90.37
UDPipe	**99.96**	50.79	95.03	95.19	92.18
NLP-Cube	99.91	84.95	58.18	95.45	92.87
RNNTagger	99.87	**99.74**	97.70	**95.95**	**93.36**
Stanza	99.92	12.75	**97.79**	**95.95**	93.20

[22] https://stanfordnlp.github.io/stanza/index.html

[23] https://github.com/dumitrescustefan/Romanian-Transformers/tree/master/evaluation

Table 4: Romanian BERTs evaluation performance on the RRT corpus for UPOS and XPOS prediction.

	UPOS Acc	**XPOS Acc**
BERT-base-ro	98.01	96.43
RoBERT-small	97.52	96.01
RoBERT-base	98.02	97.18
RoBERT-large	**98.10**	**97.65**

Regarding operations like tokenization, sentence splitting, lemmatization and part-of-speech tagging there seems to be a difference between the best tools given in-domain performances (evaluation on RRT) and the cross-domain performances (evaluation on SiMoNERo). However, RNNTagger achieves the best results for both evaluation cases for part-of-speech annotations (both UPOS and XPOS), even though Stanza achieved the same score for UPOS annotation on the SiMoNERo evaluation. Furthermore, Stanza achieved the best performance for lemmatization in both scenarios.

Unexpectedly, sentence segmentation seems to be a problem for some of the neural models when applied to the SiMoNERo corpus. This probably indicates overfitting on the training data. However, the best F1 score is achieved still by a neural model, RNNTagger, while earlier approaches achieve similar scores on the two datasets (RRT-test and SiMoNERo). One source of overfitting on the RRT corpus is the presence of sentence end characters glued to the last token, while SiMoNERo often has a space character between the last token and the sentence end character. For this reason, rule-based sentence splitters are achieving similar good performances while several neural models are facing difficulties. Given that not all tools implement dependency parsing, performance evaluation for these tasks is presented separately in Table 5, considering only the 3 tools that offer this feature: UDPipe, NLP-Cube and Stanza. Out of these, Stanza achieved the best UAS and LAS scores on both tasks.

Table 5: Performance evaluation of dependency parsing models using training on RRT and evaluation on RRT and SiMoNERo.

	RRT-test		**SiMoNERo**	
	UAS	**LAS**	**UAS**	**LAS**
UDPipe	84.35	78.64	84.01	80.24
NLP-Cube	90.65	85.87	92.07	89.09
Stanza	**92.22**	**88.08**	**92.55**	**89.51**

Training and running performance is presented in Table 6. For training we present the duration of the training process (smaller is better). When the period spans multiple hours, the number is rounded at minute level. For tokenization and overall model running we consider the number of tokens of each corpus (16324 for RRT-Test and 146020 for SiMoNERo) and give the information in tokens/second (larger numbers are better) and we round this information to an integer. These performance indicators were collected on the same system used for training, described in the "Methodology" section, above. When annotating offline corpora, annotation speed may not be a primary factor in choosing the most appropriate annotation tool. However,

it becomes increasingly important when used in (near) real-time systems, such as human-robot interactions [9]. From this perspective, data from Table 6 indicates that RNNTagger achieves the best tokenization performance, while TreeTagger the best overall speed. At the other end, NLP-Cube and Stanza achieve a lower annotation speed, given our hardware used for testing. This confirm findings by Alves et al. (2020) [2] who also indicated a high annotation time associated with NLP-Cube.

Table 6: Training duration and running speed considering tokenization and complete annotation for RRT-Test and SiMoNERo

	RRT Train	**RRT-test**		**RRT All**	**SiMoNERo**	
	Train	**Tokenize**	**Run**	**Train**	**Tokenize**	**Run**
	time	**tok/s**	**tok/s**	**time**	**tok/s**	**tok/s**
TreeTagger	70s	32648	**8162**	**21s**	43201	**24337**
TTL	**58s**	2332	583	63s	279	202
UDPipe	3h3m	9775	2720	3h15m	22024	2585
NLP-Cube	13h20m	191	60	14h54m	188	64
RNNTagger	1d3h44m	**42958**	281	1d8h46m	**73010**	453
Stanza	4d16h8m	1484	430	5d2h10m	186	44

Since we did not want to modify the code of the tools, performance measurement included also loading of saved model and word embeddings (when needed). Therefore, the strange behavior noticed when comparing the speed on the RRT-Test and SiMoNERo (where the speed is considerably higher on SiMoNERo for most tools) can be explained by the loading time of the models (the sizes are given in Table 7). With the increase of the testing corpus, the loading time of the model becomes less important in the overall time and allows for higher annotation speeds. Nevertheless, the relative order of the tools with respect to tokenization and annotation speed remains the same.

Table 7: Model sizes trained on RRT-Train and the entire RRT corpus

	RRT Train (Mb)	**RRT (Mb)**
TreeTagger	**3.8**	**4.3**
TTL	7.6	7.8
UDPipe	12.6	13.1
NLP-Cube	894.7	904.2
RNNTagger	732.8	732.8
Stanza	1005.0	1005.4

The model size presented in Table 7, takes into account all the files produced during training in the model folder. Furthermore, these files are considered without any additional compression. All the recent neural tools considered (NLP-Cube, RNNTagger and Stanza) produce larger models with a size of over 500Mb, the largest model being the one used by Stanza. Similar to the observation about running speed, the model size may be less important for usual offline corpus annotation operations. Nevertheless, the model size may be important when dealing with hardware constraints, such as those imposed by mobile or embedded systems.

5 Conclusions and Future Work

In this research we evaluated the different Romanian basic language resource kits (BLARK) offering pre-trained models. To ensure a fair comparison we re-trained the models on the Universal Dependencies version 2.7 of the RRT corpus and conducted evaluation using both same-domain data (the RRT-Test part of the corpus) and cross-domain data (the SiMoNERo corpus). As expected, recent neural models outperform older tools. However, it is not any single tool that achieves the best performance on all aspects of the Romanian language.

Considering the publication year associated with the papers describing the tools analyzed, we can draw a chart for the performance evolution of the tools. In Figure 1 and Figure 2 is presented the evolution for part-of-speech tagging, considering both UPOS and XPOS, in the same-domain (Figure 1) vs cross-domain (Figure 2) scenarios.

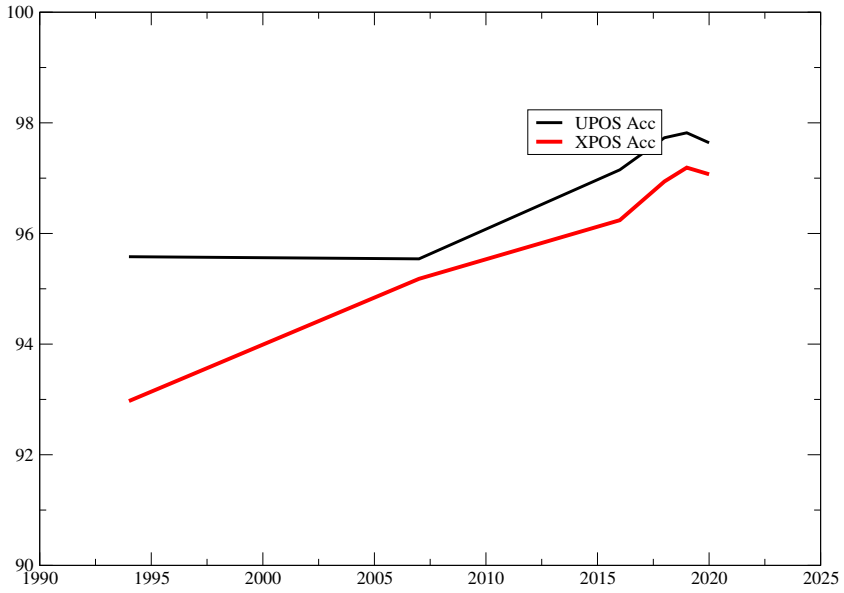

Fig. 1: Time evolution of part-of-speech annotation performance on RRT-Test

By analyzing the charts presented in Figure 1 and Figure 2 and the data from the above tables, it seems that the most recent neural approaches, such as Stanza, do not improve on the part-of-speech annotation performance (we can even notice a slight drop in this performance for certain annotations), for Romanian language, when trained in similar conditions to the other tools. Nevertheless, there is a huge increase in model size associated with recent neural tools (as presented in Table 7).

Even though the XPOS annotation performance is slightly lower than UPOS, we considered the observation from 2009 [22] suggesting that training some statistical algorithms on larger datasets could prove beneficial for predicting a smaller dataset. In theory, this could prove effective since even though the XPOS tag was wrong, the basic part-of-speech could be correct.

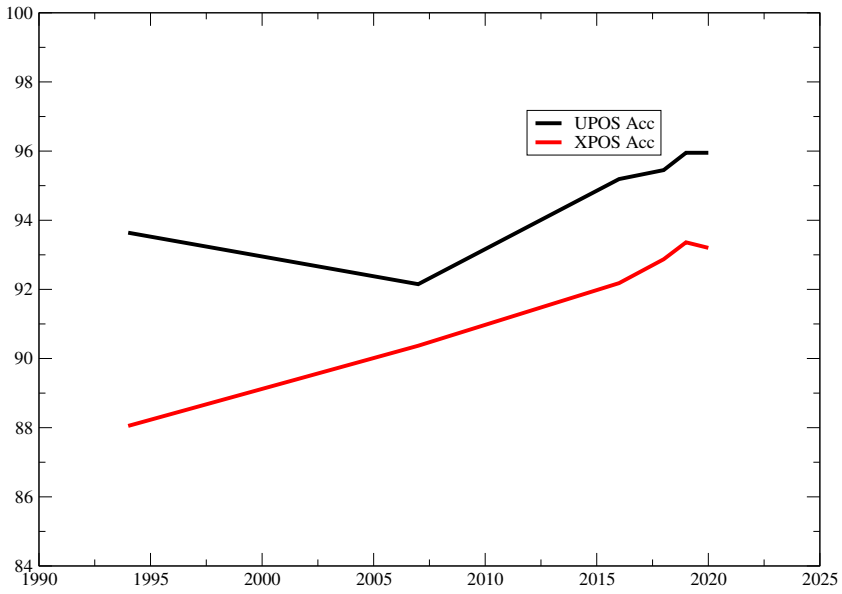

Fig. 2: Time evolution of part-of-speech tagging performance on SiMoNERo

However, our experiments indicated that this approach is no longer valid for current neural algorithms. Neither of the neural network based tools improved their results on UPOS by using the XPOS models and converting the tags. The only implementation that benefited from this approach was TTL which shows an increase in UPOS tagging accuracy from 95.54% (reported in Table 2) to 95.61% on the RRT-test corpus, when using the XPOS annotations and then converting them to UPOS.

Current state-of-the art algorithms can handle cross-domain annotations in Romanian language and the achieved results are quite high. However, there is a drop in performance for this scenario. This indicates the need for additional training data covering multiple domains. Furthermore, neural models are prone to overfitting on training data, with spectacular results on segmentation performance in the cross-domain experiment, as presented above. Recent works, such as [24], seem to indicate that noise can be used during neural network training to prevent overfitting. This is probably one way for future BLARK systems to address the issue.

Scripts and resources resulting from this work are available in a dedicated GitHub repository[24]. This allows future evaluation on different UD versions of the corpora or adding new tools as they become available. Additionally, pre-trained models for all the evaluated tools using the complete RRT corpus from UD version 2.7 are available in the RELATE[25] platform [15].

Given its recent success in all areas of natural language processing, we consider for future work creating an end-to-end system for UD parsing based on Romanian BERT, and also making a fair comparison of its performance with the performance of the tools presented in this paper on

[24] https://github.com/racai-ai/RoBLARK_evaluation
[25] https://relate.racai.ro/go/pretrainedlm

both RRT and SiMoNERo.

Acknowledgement. This research was conducted in the context of the European Language Equality (ELE) project, action ELE/101018166, work programme PPPA-LANGEQ2020.

References

[1] Adda, G., Mariani, J., Paroubek, P., Rajman, M., Lecomte, J., *L'action grace d'evaluation de l'assignation des parties du discours pour le francais*, Langues, 2(2):119–129, 1999.

[2] Alves, D., Thakkar, G., Tadić, M., *Evaluating Language Tools for Fifteen EU-official Under-resourced Languages*, in Proceedings of the 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association, pp. 1866-1873, 2020.

[3] Barbu Mititelu, V., Mitrofan, M., *The Romanian Medical Treebank - SiMoNERo*, in Verginica Barbu Mititelu, Elena Irimia, Dan Tufiș, Dan Cristea (eds), Proceedings of the 15th International Conference "Linguistic Resources and Tools for Natural Language Processing", A. I. Cuza Univesity of Iasi Publishing House, pp. 7-16, 2020.

[4] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., *Enriching Word Vectors with Subword Information*, in Transactions of the Association for Computational Linguistics, Vol 5, pp. 135-146, 2017.

[5] Boroș, T., Dumitrescu, Ș.D., Burtica, R., *NLP-Cube: End-to-End Raw Text Processing with Neural Networks*, in Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics. pp. 171-179, 2018.

[6] Brants, T., *TnT – a statistical part-of-speech tagger*, in Proceedings of the 6th Applied NLP Conference, pp. 224–231, Seattle, WA, 2000.

[7] Collins, M., *Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms*, in Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, pp. 1–8, Association for Computational Linguistics, 2002.

[8] Ion, R., *Word sense disambiguation methods applied to English and Romanian*, PhD Thesis, Romanian Academy, May 2007, 148 pages.

[9] Ion, R., Badea, V.G., Cioroiu, G., Barbu Mititelu, V., Irimia, E., Mitrofan, M., and Tufiș, D., *A Dialog Manager for Micro-Worlds*, in Studies in informatics and control, vol. 29, issue 4, 2020.

[10] Jurish, B., Würzner, K.M., *Word and sentence tokenization with Hidden Markov Models*, Journal for Language Technology and Computational Linguistics (JLCL), 28(2), pp. 61-83, 2013.

[11] Neubig, G., Dyer, C., Goldberg, Y., Matthews, A., Ammar, W., Anastasopoulos, A., Ballesteros, M., Chiang, D., Clothiaux, D., Cohn, T., *Dynet: The dynamic neural network toolkit*, arXiv:1701.03980, 2017.

[12] Mititelu, V.B., Tufiș, D., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M., *Little Strokes Fell Great Oaks. Creating Corola, The Reference Corpus of Contemporary Romanian*, Revue Roumaine de Linguistique, No./Issue 3, 2019.

[13] Paroubek, P., *Evaluating Part-of-Speech Tagging and Parsing Patrick Paroubek*, in Dybkjær L., Hemsen H., Minker W. (eds) Evaluation of Text and Speech Systems. Text, Speech and Language Technology, vol 37. Springer, Dordrecht, 2007.

[14] Păiș, V, Tufiș, D., *Computing distributed representations of words using the COROLA corpus*, Proceedings of the Romanian Academy, Series A, Vol 19, No 2, pp. 403–409, 2018.

[15] Păiș, V., Tufiș, D., Ion, R., *A Processing Platform Relating Data and Tools for Romanian Language*, in Proceedings of the 12th Language Resources and Evaluation Conference (LREC), European Language Resources Association, Marseille, France, pp. 81-88, 2020.

[16] Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D., *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*, in Association for Computational Linguistics (ACL) System Demonstrations, 2020.

[17] Schmid, H., *Probabilistic Part-of-Speech Tagging Using Decision Trees*, in Proceedings of International Conference on New Methods in Language Processing, Manchester, UK, 1994.

[18] Schmid, H., *Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts*, in Proceedings of DATeCH, Brussels, Belgium, May 2019.

[19] Spoustova, D. j., Hajič, J., Raab, J., and Spousta, M., *Semi-Supervised Training for the Averaged Perceptron POS Tagger*, in Proceedings of the 12th Conference of the European Chapter of the ACL (EACL), pp. 763–771, Athens, Greece, Association for Computational Linguistics, 2009.

[20] Straka, M., Hajič, J., Straková, J., *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), Portorož, Slovenia, 2016.

[21] Strakova, J., Straka, M., and Hajič, J., *Open-source tools for morphology, lemmatization, pos tagging and named entity recognition*, in Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 13–18, Baltimore, Maryland, Association for Computational Linguistics, 2014.

[22] Tufiş, D., *Algorithms and Data Design Issues for Basic NLP Tools*, in Language Engineering for Lesser-Studied Languages, IOS Press, pp. 3-50, 2009.

[23] Tufiş, D., Mitrofan, M., Păiş, V., Ion, R., Coman, A., *Collection and Annotation of the Romanian Legal Corpus*, in Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp. 2766-2770, 2020.

[24] You, Z., Ye, J., Li, K., Xu, Z., Wang, P., *Adversarial Noise Layer: Regularize Neural Network by Adding Noise*, in Proceedings of the IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, pp. 909-913, 2019, doi: 10.1109/ICIP.2019.8803055.

[25] Straka, M., Hajič, J., Straková, J., Hajič, J., *Parsing universal dependency treebanks using neural networks and search-based oracle*, in Proceedings of 14th International Workshop on Treebanks and Linguistic Theories (TLT 14), pp. 208-220, December 2015.

[26] Tufiş, D., Cristea, D., *IURES: A Human Engineering Approach To Natural Language Question Answering*, in W. Bibel, B.Petkoff (eds.) Artificial Intelligence: Systems, Applications, Methodology, North Holland, 1985.

[27] Estival, D., *Reversible Grammars and their Application in Machine Translation*, in Tomek Strzalkowski (ed) Reversible Grammar in Natural Language Processing, pp. 293-320, The Springer International Series in Engineering and Computer Science book series (SECS, volume 255), 1994.

[28] Estival, D., Tufiş, D., Popescu, O., *Développement d'outils et de données linguistique pour le traîtement du langage naturel*, ISSCO Public Reports, 1994.

[29] Rosner, M., Cattaneo, P., Tufiş, D., *MACPAIL: A Fruitful Cooperation between Romanian Academy and IDSIA Lugano*, in ELSNEWS vol.2, no.4, Edinburgh, 1994.

[30] Tufiş, D., *A Generalised Environment for Unification Based Natural Language Processing*, in Proceedings of the European Seminar on Language Resourses, Kaunas, April 1997.

[31] Tufiş, D., *Yet another Head Driven Generator*, in International Journal on Information and Control, vol. 3, pp. 197-208, Bucureşti, 1999.

[32] Mason, O., Tufiş, D., *Probabilistic Tagging in a Multilingual Environment: Making an English Tagger Understand Romanian*, in Proceedings of the Third International TELRI Seminar, Montecatini, October, 1997.

[33] Tufiş, D., Mason, O., *Tagging Romanian texts: A Case Study for QTAG, a Language Independent probabilistic tagger*, First International Conference on Language Resources and Evaluation, Granada, 28-30 May, 1998.

[34] Tufiş, D., *Tagging with Combined Language Models and Large Tagsets*, in Proceedings of the TELRI International Seminar on Text Corpora and Multilingual Lexicography, Bratislava, November, 1999.

[35] Tufiş, D., *Tiered tagging and combined classifiers*, in Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, pages 28–33. Springer, 1999.

[36] Dumitrescu, Ș. D., Boroș, T., Tufiș, D., *RACAI's Natural Language Processing pipeline for Universal Dependencies*, In Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Vancouver, Canada, pp. 174-181, August 2017.

[37] Păiș, V., *Multiple annotation pipelines inside the RELATE platform*, in Proceedings of the 15th International Conference on Linguistic Resources and Tools for Natural Language Processing. pp. 65-75, 2020.

[38] Păiș, V., Tufiș, D, Ion, R., *Integration of Romanian NLP tools into the RELATE platform*, in Proceedings of the International Conference on Linguistic Resources and Tools for Natural Language Processing, 2019.

[39] Barbu Mititelu, V., Mitrofan, M., *Leaving No Stone Unturned When Identifying and Classifying Verbal Multiword Expressions in the Romanian Wordnet*, in Proceedings of Global WordNet Conference. Wroclaw, Poland, pp. 10-15, jul 2019.

[40] Barbu Mititelu, V., Mitrofan, M., Mitrofan, G., *A Pilot Study for Enriching the Romanian WordNet with Medical Terms*, in Proceedings of CLiB 2018. pp. 126-134, 2018.

[41] Barbu Mititelu, V., Dumitrescu, Ș. D., Tufiș, D., *News about the Romanian Wordnet*. in Proceedings of the 7th International Global WordNet Conference, Tartu, Estonia, 2014.

[42] Tufiș, D., Barbu Mititelu, V., *The Lexical Ontology for Romanian*, in Language Production, Cognition, and the Lexicon, (Nuria Gala, Reinhard Rapp, Nuria Bel-Enguix). Springer, vol. 48, pp. 491-504, 2014.

[43] Tufiș, D., Barbu Mititelu, V., Ștefănescu, Dan., Ion, R., *The Romanian Wordnet in a Nutshell*, in Language Resources and Evaluation, vol. 47, pp. 1305-1314, 2013.

[44] Barbu Mititelu, V., *Increasing the Effectiveness of the Romanian Wordnet in NLP Applications*, in Computer Science Journal of Moldova, vol. 21, pp. 320-331, 2013.

[45] Tufiș, D., Ștefănescu, D., *Experiments with a Differential Semantics Annotation for WordNet 3.0*,in Journal on Decision Support Systems, Elsevier, pp. 695-703, 2012.

[46] Tufiș, D., Barbu Mititelu, V., Ștefănescu, D., Ion, R., *The Lexical Ontology for Romanian*, in Language Resources and Evaluation, Special Issue on Wordnets, (Calzolari, Nicoletta and Ide, Nancy), Springer, 2012.

[47] Barbu Mititelu, V., *Adding Morpho-Semantic Relations To The Romanian Wordnet*, in Proceedings of LREC2012, Istanbul, Turkey, 2012.

[48] Barbu Mititelu, V., *Wordnets: State of the Art and Perspectives. Case Study: the Romanian Wordnet*, in Proceedings of the International Conference Recent Advances in Natural Language Processing, (Angelova, G. and Bontcheva, K. and Mitkov, R. and Nikolov, N.), Hissar, Bulgaria, pp. 672-677, 2011.

[49] Barbu-Mititelu, V., Ștefănescu, D., Ceaușu, A., *Enriching the Romanian WordNet Using Semi-automatically Identified Hyponymic Patterns*, in Proceedings of the Global WordNet Conference - GWA2010, Mumbay, India, January 2010.

[50] Tufiș, D., *Paradigmatic Morphology and Subjectivity Mark-up in the RO-WordNet Lexical Ontology*, Chapter in Intelligent Systems and Technologies - Methods and Applications, (Teodorescu, H.N. and Watada, Junzo and Jain, L.), Springer Verlag, no. 217, pp. 161-179, 2009.

[51] Tufiș, D., Ion, R., Bozianu, L., Ceaușu, A., Ștefănescu, D., *Romanian Wordnet: Current State, New Applications and Prospects*, in Proceedings of the 4th Global WordNet Conference, GWC-2008, (Tanacs, Attila and Csendes, Dora and Vincze, Veronika and Fellbaum, Christiane and Vossen, Piek), Szeged, Hungary, pp. 441-452, January 2008.

[52] Tufiș, D., Ion, R., *Multilingual Word Sense Disambiguation Using Aligned Wordnets*, in ROMJIST - Romanian Journal on Information Science and Technology, vol. 7, no. 2-3, pp. 198-214, 2004.

[53] Tufiș, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L., *The Romanian Wordnet*, In ROMJIST - Romanian Journal on Information Science and Technology, vol. 7, no. 2-3, pp. 105-122, 2004.

[54] Tufiș, D., Ion, R., Ide, N., *Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets*, in Proceedings of the 20th International Conference on Computational Linguistics, COLING2004. Geneva, 2004.

[55] Tufiș, D., Barbu, A.M., Pătrașcu, V., Rotariu, G., Popescu, C., *Corpora and Corpus-Based Morpho-Lexical Processing*, Chapter in Recent Advances in Romanian Language Technology, (Tufiș, Dan and Andersen, Poul), Editura Academiei Române, București, pp. 35-56, 1997.

[56] Dimitrova, L., Ide, N., Petkevic, V., Erjavec, T., Kaalep, H. J., Tufiș, D., *Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages*, in Proceedings of the Joint 17th International Conference on Computational Linguistics 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 1998), (Boitet, Christian and Whitelock, Pete), Morgan Kaufmann Publishers, Montreal, Canada, pp. 315–319, aug 1998.

[57] Ion, R., Irimia, E., Ștefănescu, D., Tufiș, D., *ROMBAC: The Romanian Balanced Annotated Corpus*, in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), (Calzolari, Nicoletta and Choukri, Khalid and Declerck, Thierry and Uğur Doğan, Mehmet and Maegaard, Bente and Mariani, Joseph and Odijk, Jan and Piperidis, Stelios), European Language Resources Association (ELRA), Istanbul, Turkey, May 2012.

[58] Tufiș, D., Irimia, E., *RoCo_News - A Hand Validated Journalistic Corpus of Romanian*, in Proceedings of the 5th LREC Conference, Genoa, Italy, pp. 869-872, May 2006.

[59] Forăscu, C., Tufiș, D., *Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information*, in Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12), (Calzolari, Nicoletta and Choukri, Khalid and Declerck, Thierry and Uğur Doğan, Mehmet and Maegaard, Bente and Mariani, Joseph and Odijk, Jan and Piperidis, Stelios), European Language Resources Association (ELRA), Istanbul, Turkey, May 2012.

[60] Irimia, E., Barbu Mititelu, V., *Building a Romanian Dependency Treebank*, in Corpus Linguistics 2015, Lancaster University, UK, 2015.

[61] Barbu Mititelu, V., Mărănduc, C., Irimia, E., *Universal and Language-specific Dependency Relations for Analysing Romanian*, in Proceedings of DepLing2015, 2015.

[62] Mitrofan, M., Barbu Mititelu, V., Mitrofan, G., *MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language*, in Proceedings of the BioNLP workshop, Association for Computational Linguistics, Florence, Italy, pp. 71-79, aug 2019.

[63] Vlad, A., Mitrea, A., Ciucă, Ș., Luca, A., *A Study on the Statistical Structure of Words and of Word Digrams in a Literary Romanian Corpus*, in Proceedings of the 6th International Conference Speech Technology and Human-Computer Dialogue "SpeD 2011", Brașov, România, pp. 151-158, May 2011.

[64] Tufiș, D., Ceaușu, A., Ion, R., Ștefănescu, D., *An integrated platform for high-accuracy word alignment*, in Proceedings of JRC Enlargement and Integration Workshop: Exploiting parallel corpora in up to 20 languages, Arona, Italy, September 2005.

[65] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiș, D., *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages*, in Proceedings of the 5th LREC Conference, Genoa, Italy, pp. 2142-2147, May 2006.

[66] Ceaușu, A., *Colectarea și procesarea documentelor românești ale corpusului JRC-Acquis*, in Atelierul de Resurse Lingvistice pentru Limba Română (CONSILR 2008), Editura Universității "Al. I. Cuza", Iași, November 2008.

[67] Skadina, I., Vasiljevs, A., Skadins, R., Gaizauskas, R., Tufiș, D., Gornostay, T., *Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation*, in Proceedings of the 3rd Workshop on Building and Using Comparable Corpora (BUCC) at the 7th Language Resources and Evaluation Conference (LREC 2010), Valetta, Malta, pp. 6-14, May 2010.

[68] Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pezik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., Brank, J., *The MARCELL Legislative Corpus*, in Proceedings of The 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp. 3754-3761, May 2020.

[69] Tufiș, D., Păiș, V., Barbu Mititelu, V., Ion, R., Irimia, E., Avram, A., Curea, E., *MARCELL – A project to remember: hard work of a friendly consortium under wise coordination*, Chapter in A korpusznyelvészettől a neurális hálókig : Köszöntő kötet Váradi Tamás 70, születésnapjára (Dodé, Réka and Ludányi, Zsófia), Nyelvtudományi Kutatóközpont, Budapest, pp. 127-140, 2021.

[70] Tufiș, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ș. D., Boroș, T., *The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language*, in Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pp. 2516–2521, May 2016.

[71] Tufiș, D., Barbu Mititelu, V., Irimia, E., Dumitrescu, Ș. D., Boroș, T., Teodorescu, N., Cristea, D., Scutelnicu, A., Bolea, C., Moruz, A., Pistol, L., *CoRoLa Starts Blooming – An Update on the Reference Corpus of Contemporary Romanian Language*, in Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), 2015.

[72] Barbu Mititelu, V., Irimia, E., Tufiș, D., *CoRoLa – The Reference Corpus of Contemporary Romanian Language*, in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, pp. 1235–1239, 2014.

[73] Barbu Mititelu, V., Cosma, R., Cristea, D., *Corpus of Contemporary Romanian. Architecture, Annotation Levels and Analysis Tools*, in Lingvistică românească, lingvistică romanică. Actele celui de-al XVI-lea Colocviu Internațional al Departamentului de Lingvistică (Helga Bogdan Oprea, Andreea-Victoria Grigore, Rodica Zafiu), University of Bucharest Publishing House, pp. 13-20, 2017.

[74] Barbu Mititelu, V., Tufiș, D., Irimia E., *The Reference Corpus of the Contemporary Romanian Language (CoRoLa)*, in Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 1178-1185, 2018.

[75] Kupietz, M., Diewald, N., Trawiński, B., Cosma, R., Cristea, D., Tufiș, D., Váradi, T., Wöllstein, A., *Recent developments in the European Reference Corpus (EuReCo)*, in Louvain-la-Neuve: Presses universitaires de Louvain, pp. 257-273, 2020.

[76] Păiș, V., Ion, R., Barbu Mititelu, V., Irimia, E., Mitrofan, M., Avram, A., *ROBIN Technical Acquisition Speech Corpus*, Zenodo, doi:10.5281/zenodo.4626539, mar 2021.

[77] Boroș, T., Dumitrescu, Ș. D., Păiș, V., *Tools and resources for Romanian text-to-speech and speech-to-text applications*, in CoRR. vol. abs/1802.05583, 2018.

[78] Dumitrescu, Ș.D., Boroș, T., Ion, R., *Crowd-sourced, automatic speech-corpora collection – building the Romanian Anonymous Speech Corpus*, in Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL2014), Reykjavik, Iceland, May 2014.

[79] Boroș, T., Stan, A., Watts, O., Dumitrescu, Ș.D., *RSS-TOBI - A Prosodically Enhanced Romanian Speech Corpus*, in Language Resources and Evaluation Conference (LREC 14), Reykjavik, Iceland, May 2014.

[80] Cristea,D., Pistol, I.,Boghiu, Ș., Bibiri, A.D., Gîfu, D., Scutelnicu, A., Onofrei, M., Trandabăț, D., Bugeag, G., *CoBiLiRo: A Research Platform for Bimodal Corpora*, in Proceedings of the 1st International Workshop on Language Technology Platforms, May 2020.

[81] Barbu Mititelu, V., Irimia, E., Păiș, V., Avram, A.M., Mitrofan, M., Curea, E., *Romanian Resources in LLOD format*, in Proceedings of the 15th International Conference Linguistic Resources and Tools for Natural Language Processing, (Barbu Mititelu, Verginica and Irimia, Elena and Tufiș, Dan and Dan, Cristea), pp. 29-40, December 2020.

[82] Dumitrescu, S., Avram, A.M. and Pyysalo, S., *The birth of Romanian BERT*, In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 4324-4328, 2020.

[83] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I.,*Attention is all you need*, arXiv preprint arXiv:1706.03762, 2017.

[84] Masala, M., Ruseti, S. and Dascalu, M., *RoBERT–A Romanian BERT Model*, In Proceedings of the 28th International Conference on Computational Linguistics, pp. 6626-6637, 2020.

[85] Păiș, V., Mitrofan, M., Gasan, C.L., Ianov, A., Ghiță, C., Coneschi, V.S., Onuț, A., *Romanian Named Entity Recognition in the Legal domain (LegalNERo)*, Zenodo, doi: 10.5281/zenodo.4772094, 2021.

[86] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arXiv:1810.04805 [cs.CL], 2019.

[87] Barbu Mititelu, V., *Modern syntactic analysis of Romanian*, In Ichim, O. (coord.), Botoşineanu, L., Butnaru, D., Clim, M.R., Ichim, O., Olariu, V. (eds.), Clasic şi modern în cercetarea filologică românească actuală, Iaşi, Editura Universităţii "Alexandru Ioan Cuza", pp. 67-78, 2018.