

A presentation of the Research Institute for Artificial Intelligence "Mihai Drăgănescu"

Welcome to HE ambassador of  
Republic of India, Mr. Rahul  
Shrivastava



# The plan of the meeting

- Part I

- 14:10-14:30 Acad. Dan Tufis: Presentation of the Institute:
  - Domains of interest and activities,
  - Main achievements
  - International positioning of the institute
- 14:30-14:40 Dr. Verginica Barbu Mititelu: Innovation in language resources development
  - Major resources developed
  - Standardization (open linked data format)
  - Access to the resources
- 14:40-14:50 Acad. Dan Tufis: Open portal for technologies for the Romanian language
  - Main characteristics of the portal, services and access
  - Availability

- PART II

- 9:50 - Question answering, Open discussions

# Who are we (ICIA)?

- A small research institute, established in 2002 on the premises of a smaller Center for Advanced Research established in 1994
- Permanent personnel: research positions 18, administrative: 6
- A larger number of internships (more than 25 every year)
- PhD programs, small number of PhD students, only 3 PhD advisors
- Organizer/co-organizer of various international scientific & educational events (15 editions of the EUROLAN intl. Training school, 15 editions of CONSILR conference, 11 editions of SPeD, 3 ELRC workshops, etc)
- Solid publications record (more than 1070 publications)
- Large scientific partnerships ( more than 50 intl projects and 34 natl. Projects)
- ISP for most of the institutes in the House of the Academy
- A decent research infrastructure (10 servers both CPU and GPU equipped, individual desktops and laptops/tablets, smart router, several switches, etc)

# A few remarkable results of ICIAR researchers

- The construction and maintenance of one of the world largest lexical resources: RoWordNet (more than 20 years)
- Preparing the Romanian part of the JRC-Aquis multilingual corpus (the most used multilingual resource in training MT systems)
- Aligning large comparable corpora and extracting parallel text segments (ACCURAT project)
- Implementing the first professional MT systems for Romanian in various technological epochs (dictionary-word based, grammar rule-based, statistics-based, deep NN-based)
- Winning a series of shared-tasks (intl. competitions):
  - word alignment (ACL, Edmonton, 2003), word-sense disambiguation (ACL SenseEVAL III, Barcelona, 2004), word alignment (Ann Arbor (ACL, 2005), CLEF Q&A (2007, 2008, 2009, 2010), several other shared tasks in which our team was among the protagonists NER New Mexico (ACL, 2021), Punta Cana (ACL, 2021) etc.

**Data** is the  
new **oil** of the  
21st century



## Data is the new oil.

We see in data the same transformative, wealth-creating power that 19th-century visionaries once sensed in the crude black ooze trapped underground.

If "crude" data can be extracted, refined, and piped to where it can impact decisions in real time, its value will soar. And if data can be properly shared across an entire ecosystem and made accessible in the places where analytics are most useful, then it will become a true game changer, altering the way we live, work, learn, and play.



Source: Cisco IBSG, 2012. #DataInMotion

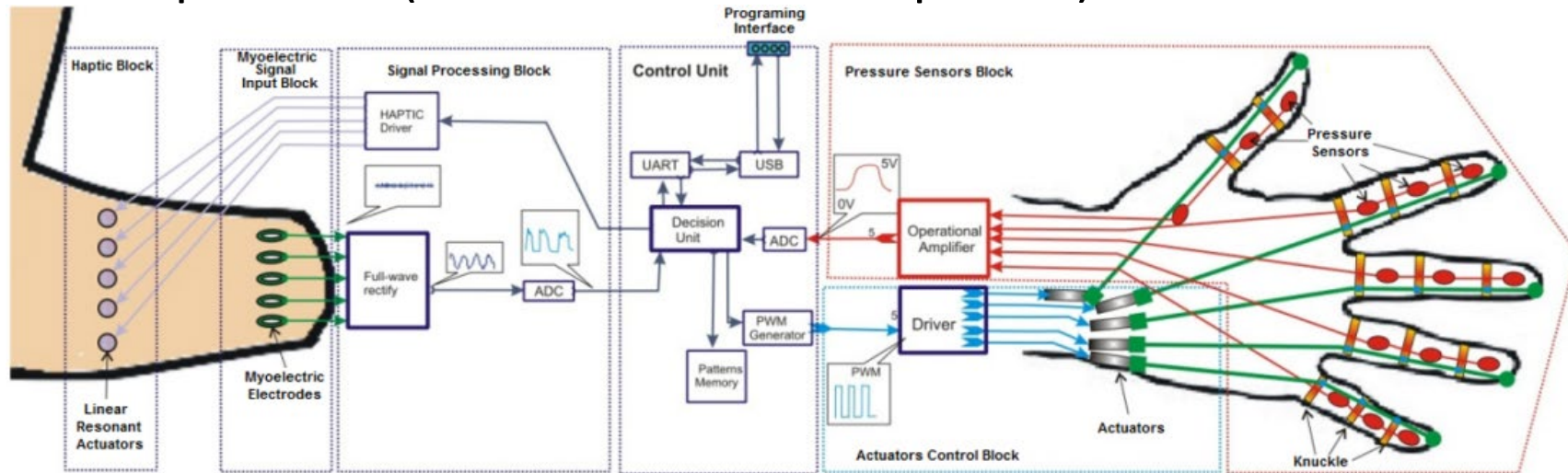


# Research and Development Areas

- Large language Resources and development of SOTA Technologies
  - Language resources in a standardized reusable form
  - Instruments for processing large quantities of language data
  - Neural networks language models (BERT-like)
  - Applications mediated by language (Romanian) (QA, MT, chatbots, tutoring systems, etc.)
- Language Technology infrastructures
  - Trans-European Language Resources Infrastructure, European Network of Excellence in Human Language Technologies, Fostering Language Resources Network, Enhancing the European Linguistic Infrastructure, ): European network for Web-centred linguistic data science (LLOD), European Language Grid, European Language Equality , European Language Research Coordination, Portal of Romanian Language Technology (tools and data)

# Research and Development Areas

- National Center for the Study of the Brain (newly established), coordinated by 2 members of the Romanian Academy and a distinguished Professor of neuro sciences
- New Electronic Architectures and applications
  - Myoelectric prosthesis (artificial arms for amputees)



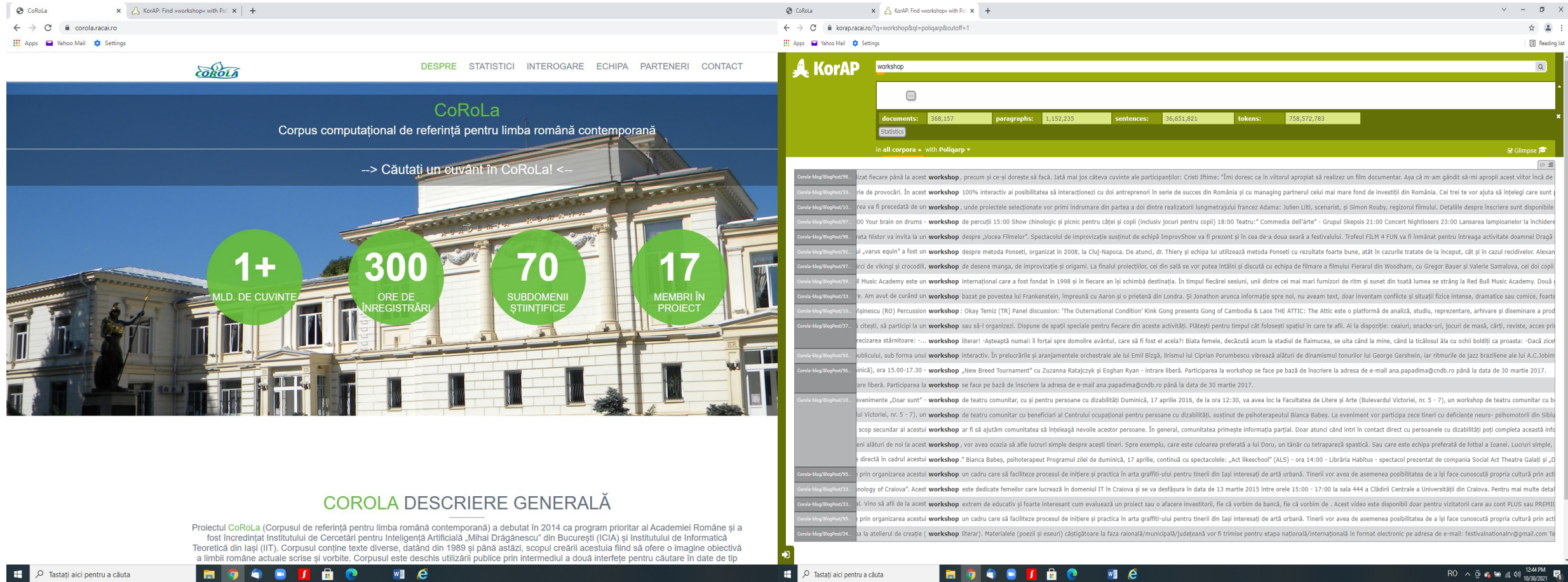
- Study of the babies cry (Dunstan language – universal?)
- Study and combatting the Fake News (a recent Horizon-IA project proposal led by the governmental *Department of Relations with Moldovan Republic*)



Major results shared with everybody (I)

The Reference Corpus of Contemporary Romanian Language

ICIA, UB, IIT, UAIC, IDS-Mannheim





# Major results shared with everybody (II)

## Portalul RELATE

RELATE

relate.racai.ro

Apps | Yahoo Mail | Settings

Reading list

RELATE

Romanian Portal of Language Technologies

Login

TEPROLIN Service >

CoRoLa >

RoWordNet >

Machine Translation >

Speech >

Named Entity Recognition

CURLICAT Anonymization

EUROVOC Classification >

Punctuation Restoration >

Pretrained LM >

Citation >

JSON | CoNLL-U | CoNLL-X | XML | Text | Chunks | Tree | Entities

Fiscul vă face verificări la firmele indicate de CNSP, iar pe zona de dezvoltare va acorda granturi, precum cele pentru primarii.

Word	acorda
Lemma	acorda
U-POS	VERB
CTAG	VN
MSD	Vmnp
Chunk	Vp#3
Named Entity	
Phonetic	a k o r d a
Syllables	a-cor-'da
Similar Words	acordă acordat acordată primi beneficia acorde

Tastați aici pentru a căuta

1:06 PM 10/30/2021

# Major results shared with everybody (III)

## ICIA github

https://github.com/racai-ai

RACAI - GitHub

- pyeurovoc** Public  
Legal document classification with EuroVoc descriptors on 22 languages.  
Python 0 0 0 0 Updated Aug 26, 2021
- LegalNER** Public  
NER in the Legal domain  
Java 0 0 0 0 Updated May 14, 2021
- Romanian-DistilBERT** Public  
This repository contains the Romanian version of DistilBERT.  
Jupyter Notebook 0 0 0 0 Updated May 5, 2021
- ROBINDialog** Public  
This is the micro-world dialog manager developed in the ROBIN project.  
Java 0 0 0 0 Updated Apr 20, 2021
- TEPROLIN** Public  
This is the TEPROLIN Romanian text processing platform, developed in the ReTeRom project.

Tastați aici pentru a căuta

https://github.com/racai-ai

RACAI - GitHub

Find a repository...

RELATE Public  
RELATE platform for processing Romanian language  
JavaScript 1 0 0 1 Updated Oct 26, 2021

Rodna Public  
Romanian Deep Neural Network Architectures project  
Python 0 0 0 0 Updated Oct 26, 2021

ROAnonymization\_CURLICAT Public  
Romanian text anonymization (pseudonymization) from the CURLICAT project  
Java 0 0 0 0 Updated Oct 20, 2021

PunctuationRestoration Public  
Experiments with punctuation restoration  
Python 0 0 0 0 Updated Oct 5, 2021

ro-ud-autocorrect Public

https://github.com/racai-ai

RACAI  
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy  
Bucharest, Romania <http://www.racai.ro/en/> Verified

Overview Repositories Packages People Projects

Popular repositories

- pyeurovoc** Public  
Legal document classification with EuroVoc descriptors on 22 languages.  
Python 7 1
- RELATE** Public  
RELATE platform for processing Romanian language  
JavaScript 1
- IATE-EUROVOC-Annotator** Public
- RobinASR** Public  
Romanian Automatic Speech Recognition from the ROBIN project  
Python 2 4
- RoLLOD** Public  
Tools for Romanian Linguistic Linked Open Data  
Java 1
- TermEval2020** Public  
Automatic Term Extraction (ATE) system that participated in the TermEval 2020 competition

People  
This organization must be a member organization.

Top languages  
Loading...

Most used top  
Loading...

Tastați aici pentru a căuta

12:49 PM 10/30/2021

Thank you!