Romanian Schools of Computational Linguistics and Natural Language Processing Technologies and their Achievements

Dan Cristea
“Alexandru Ioan Cuza” University of Iași,
Faculty of Computer Science
Iași branch of the Romanian Academy,
Institute of Computer Science
dcristea@info.uaic.ro
Speech & Dialogue Research Laboratory

• Host: University Politehnica of Bucharest, Faculty of Electronics, Telecommunications and Information Technology
• URL: [https://speed.pub.ro/](https://speed.pub.ro/)
• Leader: Corneliu Burileanu (since 1984)
• Members: 4 academics, 4 PhD, 4 PhD stud., master&batchelor stud.
• Areas: automatic speech recognition, Text-to-Speech synthesis, speaker recognition, spoken term detection, spoken document indexing/retrieval, language modeling, digital signal processing, etc.
• Approaches: HMM, statistical
• Projects (UEFISCDI, Romanian-American Foundation): Voice-controlled Assistive System for Intelligent Buildings (ANVSIB), Phonetic Analysis of the Romanian Language (AFLR), Automatic Baby-Language Recognition System (SPLANN), Enhanced Large Vocabulary Continuous Speech Recognition (LVCSR) for Romanian, etc.

• Applications: Rich Speech Transcription Service, SpeeD Dictation, etc.

• School: 10 past PhDs & 4 PhD students (prof. Burileanu), Master in Multimedia Technologies in Biometrics and Information Security Applications (BIOSINF)
SpeD series of conferences (since 2001)

In collaboration with ARFI-IIT, etc.
Speech-Group@ARFI-IIT

- **Host:**
  - Iași branch of the Romanian Academy, Institute for Computer Science
- **Leader:** H.N. Teodorescu
Speech Processing Group
@Tech. University Cluj

- Host: Technical University of Cluj-Napoca, Communications Department
- URL: [http://speech.utcluj.ro/](http://speech.utcluj.ro/)
- Leader: Mircea Giurgiu
- Members: 4 PhD and 1 PhD stud., 4 alumni
- Areas: speech resources and dictionaries, speech segmentation, speech-to-text alignment, speech recognition and synthesis, lexical stress detection, prosody and intonation, eliminating noise in speech, etc.
- Approaches: HMM, neural, lightly supervised learning, etc.
Speech Processing Group
@Tech. University Cluj

- Projects: SWARA (PN II), SIMPLE4ALL (FP7), Sound2Sense (Marie Curie)
- Resources: RSS-TOBI – Romanian prosody corpus (in collaboration with RACAI), MaRePhoR – machine-readable phonetic dictionary for Romanian, SWARA – parallel Romanian read speech dataset, RSS – Romanian speech synthesis corpus
- School: PhD, Master in Multimedia Technologies
NLP-Group@RACAI, Romanian Academy

- Leader: Acad. Dan Tufiş
Human Language Technologies Research Center

- Host: Faculty of Mathematics and Computer Science, University of Bucharest
- URL: http://nlp.unibuc.ro/
- Leaders: Solomon Marcus (past), Liviu P. Dinu
- Members: 7 academics & PhD, 8 PhD studs, 9 ext. collaborators
- Areas: digital humanities, phonological studies, formal and distributional semantics, authorship identification, plagiarism detection, WordNet, WSD in IR, etc.
- Approaches: formal, quantitative (including machine learning), multi-criteria methods
Human Language Technologies Research Center

- Projects financed by: UEFISCDI (PNCDI III, PN II Idei, PNCDI2), CNCSIS, MedC-ANCS, Univ. Bucharest;
- Resources: corpus of native, non-native and translated texts, dataset of speech and text features, contribution to English WordNet (relation instance-of), Romanian WSD task in SENSEVAL-3
- Tools: neural text simplification toolkit, Romanian accent detection, author identification etc.
- School: many past PhD (S.Marcus), ongoing: 7 (Dinu)
- Events:
  - *Solomon Marcus* Seminar of Mathematical and Computational Linguistics
  - *Recent Advances in Artificial Intelligence Conference* – 1st edition 2017
K-Teams
Computer Supported Collaborative Knowledge Construction Laboratory

• Host: University Politehnica of Bucharest, Faculty of Automatic Control and Computers

• URL: https://cs.pub.ro/index.php/83-laboratories/274-k-teams-computer-supported-collaborative-knowledge-construction-laboratory

• Leader: Ştefan Trăşan-Matu

• Members: 7 academics, more PhD studs.

• Areas: written text, discourse analysis (chats), sentiment mining, intertextuality, opinion propagation in social networks, semantic web, etc.

• Approaches: statistical, neural
K-Teams

- **Projects:** CNCSIS, COST, NSF, EUREKA, FP7, H2020, etc.
- **Resources:** corpus of online conversations (chat), 63M words corpus
- **Tools:** PolyCAFe – chat analysis platform, ReaderBench – data mining and discourse analysis multi-purpose platform for Romanian, English, French, Dutch and Latin
- **School:** 12 past PhD & 11 ongoing (Trăușan-Matu)
- **Events:**
  - **DS-CSCL** - International Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning
  - **RoCHI** - Romanian Conference on Human-Computer Interaction
NLP-Group@UAIC-FII
NLP-Group@ARFI-IIT

• Hosts:
  – “Alexandru Ioan Cuza” University of Iași, Faculty of Computer Science
  – Iași branch of the Romanian Academy, Institute for Computer Science

• URL: http://nlptools.info.uaic.ro/

• Leader: Dan Cristea

• Members: 15+ PhD, 5 PhD stud., 25+ alumni

• Areas: written text, building resources, segmentation, morphology, syntax, discourse structure, anaphora resolution, summarisation, sentiment analysis, etc.

• Approaches: symbolic (rule-based), statistical
Projects: FP6, FP7, COST, CNCSIS, UEFISCDI, Univ. Iași

Resources: Ro-WN (in coll.), eDTLR (in coll.), COROLA (in coll.), UAIC-RoTB, RoFrameNet, QuoVadis, etc.

Tools:
- POS-tagger, lemmatiser, GGS – Graphical Grammar Studio, NP-chunker, FDG syntactic parser, anaphora resolver, clause splitter, discourse parser, summariser, dictionary entry parser,

School: 7 past PhD & 5 ongoing (prof. Cristea), Master in Computational Linguistics

Events: BringITon! Series of workshops (since 2010), ...
In collaboration with RACAI, etc.

EUROLAN series (since 1993)
WELCOME TO EUROLAN 2017

SUMMER SCHOOL ON BIOMEDICAL TEXT PROCESSING
10 - 17 SEPT. 2017, CONSTANȚA, ROMANIA
THE 13TH EUROLAN SCHOOL ON NATURAL LANGUAGE PROCESSING
Annual series of conferences
“Linguistic Resources and Tools for Romanian Language”
(since 2001)

In collaboration with RACAI, etc.
Other centers

• Statistical approaches on Romanian language – prof. Adriana Vlad – University POLITEHNICA Bucharest

  – computational semantics, formal and quantitative approaches
  – impressive alumni (Daniel Marcu, Rada Mihalcea, Constantin Orasan, etc.)

• Former group of Prof. Țăndăreanu, Univ. of Craiova
  – AI techniques applied in NLP
  – 2 PhD in NLP (one finalised at West Univ. Timișoara)

• Evolving interest at West Univ. Timișoara (2 PhD in progress – prof. V.Negru)
Conclusions

• Rather rich traditions
  – ML: years ’60, S.Marcus
  – NLP: years 80’, first QA systems built

• Considerable NLP diaspora
  – “Why are so many Romanians in the field of CL?” – at CiCLing-2010 (Iași)

• Today: progress made since the 2012 META-NET White Paper (Springer)
  – BUT: lack of dedicated financing => many talented young people are lost
Thank you!