C.1 CURRICULUM VITAE

Personal information

Name, Surname:	Dan Tufiș		
Date of birth:	5 February 1954	Sex:	male
Nationality:	Romanian		
Researcher unique identifier(s) (ORCID, Researcher ID etc.):	https://orcid.org/0000-0002-8280-9852, brain-map ID: U-1700-027P-6123		
URL for personal website (if case):	https://www.racai.ro/about-us/racai-staff/tufis/		

Education

Year	Faculty of Computer Science and Automatization/Department Computer Science- University ”Politehnica” of Bucharest - Romania
1992 (dissertation defended)	Ph.D. ("Mediu de dezvoltare a sistemelor de dialog în limbaj natural")
1979	Master

Positions - current and previous
(*Academic sector/research institutes/industrial sector/public sector/other*)

actualmente august 2002 – mai 2024	Senior Researcher I, Director de Onoare Director ICIA – Research Institute for Artificial Intelligence „Mihai Drăgănescu" of the Romanian Academy http:// www.racai.ro
decembrie 1998- iulie 2002	Director CCAIAPLNMC – Romanian Academy
septembrie 1989 - decembrie 1998	Head of Laboratory ICI – National Institute for Research and Development in Informatics
septembrie 1980 – septembrie 1989	Research Group Lead ICI – Institutul Național de Cercetare Dezvoltare în Informatică

Career breaks (if case)

Year	Reason
yyyy-yyyy	Not the case

Project management experience
(*Academic sector/research institutes/industrial sector/public sector/other. Please list the most relevant.*)

1. "European Language Resource Coordination - ELRC" (2015-2021), 30 de parteneri. Membru al Language Resource Board (LRB) organismul de conducere al ELRC
2. "ELG – European Language Grid" (Horizon 2020, 2019-2022, agreement no.825627) – I am representing the Romanian Competence Center
3. "CURLICAT - Curated Multilingual Language Resources for CEF AT" (INEA/CEF/ICT/A2019/1926831), 2020-2022
4. "CEF Automated Translation Toolkit for the EU Council Presidency" (INEA/CEF/ICT/A2018/1762093) 2018-2020
5. " MARCELL - Multilingual Resources for CEF.AT in the legal domain" (INEA/CEF/ICT/A2017/1565710), 2018-2021
6. "NexusLinguarum – European Network for Research and Development in Linguistic Data Science" (OC-2018-2-23228) 2019-2022

7. PARSEME (Parsing Multiword Expressions) –COST Action IC1207 (2013-2017), 30 countries, 29 official languages and 6 dialects
8. Interacademic project (Bulgaria-România) " Enhanced Knowledge Bases for Bulgarian and Romanian" (2015-2017).
9. DRuKoLA – Project funded by Alexander von Humbold Fundation, in partnership with IDS-Mannheim, University of București and ITI-Iași (2016-2018)
10. ReTeRom - Resurse și tehnologii pentru dezvoltarea interfețelorom-mașina în limba română PN-III-P1-1.2-PCCDI-2018-0818, nr. 73PCCDI/2018, 2018-2020
11. ROBIN – Roboții și Societatea: Sisteme Cognitive pentru Roboți Personali și Vehicule Autonome PN-III-P1-1.2-PCCDI-2018-0818, nr. 72PCCDI/2018, 2018-2020
12. CoRoLa –Prioritary Program of the Romanian Academy 2017-2019

Other relevant professional experiences
(e.g. institutional responsibilities, organisation of scientific meetings, membership in academic societies, review boards, advisory boards, committees and major research or innovation collaborations, other commissions of trust in public or private sector)

Full Member of the Romanian Academy (Academician)
Awarded by the President of Romania the Order "Steaua României" (Knight degree)
Doctor Honoris Causa of the University Agora of Oradea
Honorary Professor of the University "A.I. Cuza" din Iași
The Romanian Academy prize „Traian Vuia" for research in the area of computational morphology (paradigmatic morphology and the PARADIGM system)
The Romanian Academy prize „Tudor Tănăsescu" for research in the area of linguistic modelling based on unification (MAC-ELU system)
Merit Diplomas for essential contributions to the research programs at the National Institute for Research and Development in Informatics and Romanian Academy.

More than 4450 de citations, h-index 30, i10-index 88

PhD Advisor&Director (since 2003)

Member in professional organisations:
- Professional Member of ACM

- European Language Grid (National Competence Center)
- European Language Resource Coordination (National Technological Contact Point, member of the board LRB)
- European Association for Computational Linguistics **EACL**,
- European Association for Machine Translation**, EAMT**
- Expert Advisory Group on Language Engineering Standards **EAGLES**,
- Trans European Language Resources Infrastructure **TELRI Gmb**,
- Association for Computational Linguistics Special Interes Group on Lexicons **ACL-SIGLEX**,
- Global WordNet Association **GWA**,
- Forumul Român pentru Societatea Cunoașterii (founding member)
- Asociația Română de inteligență Artificiala -**ARIA** (founding member)
- Asociația Română de Lingvistică Computațională – **ARLC** (founding member)
- Comisia Academiei Române de Informatizare pentru Limba Română (President)

Member of the Editorial Boards of several international journals:
- International Journal on Language Resources and Evaluation, Springer Verlag, USA (from 2002)
- The Computer Science Journal of Moldova, Republica Moldova (from 2005)
- Proceedings of the Romanian Academy, Series A (din 1999)- topical editor
- Romanian Journal on Science and Technology of Information, Academia Română (from 1997)
- International Journal on Information and Control, ICI, România (from 1992)
- International Journal of Computers, Communication & Control, Universitatea „Agora" din Oradea (from 2006)
- Romanian Journal of Human – Computer Interaction, ACM SIGCHI Romania (from 2008)
- NOESIS – Travaux du comite roumain d'histoire et de philosophie de sciences, Academia Româna (from 2007) –director (2015-2020)

Member of more than 150 de Program Committees for international conferences

- Member in more than 60 international și national PhD awarding committes
- University „A. I. Cuza" of Iași
- University "Politehnica" of București,
- University București,
- University of Craiova,
- Academia Română,
- University "Luis Pasteur" of Strasbourg, France
- University of Science and Technologies of Lille, France
- University of Pisa, Italy
- Mumbay Institute de Tehnology in India,
- Institutul de Matematică și Informatică al Academiei Moldovei in Chișinău
- Politechnic Institute in Grenoble, France
- Universitatea Tehnică din Strasbourg, France
- University in Alicante, Spain

C.2 Track record of the last 2 years

1. Păiş, Vasile and Tufiş, Dan and Irimia, Elena and Barbu Mititelu, Verginica. **Building a corpus for the anonymization of Romanian jurisprudence.** In *Proceedings of The 18th Linguistic Annotation*

Workshop (LAW-XVIII) (Henning, Sophie and Stede, Manfred). Association for Computational Linguistics, St. Julians, Malta, pp. 71--76, mar 2024 **[bib]** **[link]**

2. Păiș, Vasile and Barbu Mititelu, Verginica and Irimia, Elena and Ion, Radu and Badea, Valentin and Tufiș, Dan and Sterea-Grossu, Cosmin. **Towards anonymizing the Romanian jurisprudence.** In *The 18th International Conference on Linguistic Resources and Tools for Natural Language Processing (ConsILR-2023)*. 2023 **[bib]**

3. Păiș, Vasile and Tufiș, Dan. **Language Report Romanian.** In *European Language Equality: A Strategic Agenda for Digital Language Equality* (Rehm, Georg and Way, Andy). Springer International Publishing, Cham, pp. 199--202, 2023 **[bib]** **[link]**

4. Avram, Andrei-Marius and Smădu, Răzvan-Alexandru and Păiș, Vasile and Cercel, Dumitru-Clementin and Ion, Radu and Tufiș, Dan. **Towards Improving the Performance of Pre-Trained Speech Models for Low-Resource Languages Through Lateral Inhibition.** In *2023 46th International Conference on Telecommunications and Signal Processing (TSP)*. vol. , no. , pp. 234-237, 2023 **[bib]**

5. Nicolae, Dragoș, Constantin and Yadav, Rohan, Kumar and Tufiș, Dan. **Evaluation of Language Models on Romanian XQuAD and RoITD datasets.** In *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*. vol. 18, no. 1, 2023 **[bib]** **[link]**

6. Avram, Andrei-Marius and Catrina, Darius and Cercel, Dumitru-Clementin and Dascalu, Mihai and Rebedea, Traian and Păiș, Vasile and Tufiș, Dan. **Distilling the Knowledge of Romanian BERTs Using Multiple Teachers.** In *Proceedings of the Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pp. 374--384, June 2022 **[bib]** **[link]**

7. Nicolae, Dragoș, Constantin and Yadav, Rohan, Kumar and Tufiș, Dan. **A Light Romanian BERT: ALR-BERT.** In *Computers*. vol. 11(4), no. 57, 2022 **[bib]**

8. Păiș, Vasile and Avram, Andrei-Marius and Ion, Radu and Tufiș, Dan. **The Impact of Speech Recognition Performance on Human-Computer Interaction.** Chapter in *Speech Recognition Technology and Applications* (Păiș, Vasile). Nova Science Publishers, pp. 57--84, 2022 **[bib]** **[link]**

9. Avram, Andrei-Marius and Păiș, Vasile and Tufiș, Dan. **Terminology Annotation Using a String-Matching Algorithm.** In *Proceedings of the 17th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing (CONSILR 2022)*. Chișinău, Moldova, 2022 **[bib]** **[link]**

10. Váradi, Tamás and Tadić, Marko and Koeva, Svetla and Ogrodniczuk, Maciej and Tufiș, Dan and Garabík, Radovan and Krek, Simon and Repar, Andraž. **Curated Multilingual Language Resources for {CEF} {AT} ({CURLICAT}): overall view.** In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. European Association for Machine Translation, Ghent, Belgium, pp. 339--340, june 2022 **[bib]** **[link]**

11. Avram, Andrei-Marius and Păiș, Vasile and Tufiș, Dan. **Self-Supervised Pre-Training in Speech Recognition Systems.** Chapter in *Speech Recognition Technology and Applications* (Păiș, Vasile). Nova Science Publishers, pp. 27--56, 2022 **[bib]** **[link]**

12. Păiș, Vasile and Tufiș, Dan. **Capitalization and punctuation restoration: a survey.** In *Artificial Intelligence Review*. vol. 55, no. 3, pp. 1681-1722, 2022 **[bib]** **[link]**

13. Avram, Andrei-Marius and Păiș, Vasile and Tufiș, Dan. **Terminology Annotation Using a String-Matching Algorithm.** In *Proceedings of the 17th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing (CONSILR 2022)*. Chișinău, Moldova, 2022 **[bib]** **[link]**

14. Dan Tufiș. **Romanian Language Technology – a view from an academic perspective.** In *INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL*. vol. 17, no. 1, 2022 **[bib]** **[link]**

C.3 Narrative CV

Research interests

- Artificial Intelligence, Large Language Models
- Natural Language Processing
- Machine Translation
- Language Technologies
- Machine Learning
- Natural Language Understanding and Generation
- Knowledge Representation
- Semantic Web

International Research projects

- **LLMs4EU -** Call: DIGITAL-2024-AI-B-06 (Cloud Data and Artificial Intelligence) Topic: DIGITAL-2024-AI-B-06-LANGUAGE, Type of Action: DIGITAL-SIMPLE (DIGITAL Simple Grants) Proposal number: 101198470
- SAROJ
- CURLICAT – Curated Multilingual Language Resources for CEF AT (INEA/CEF/ICT/A2019/1926831), 2020-2022
- ELRC – European Language Resource Coordination – ELRC (CEF SMART 2014/1074), (2015-2021)
- ELG- European Language Grid" (Horizon 2020, agreement no.825627) 2019-2022
- ELE- European Language Equality (PPPA-Langeq-2020) 2021-2022
- Nexus Linguarum – European Network for Research and Development in Linguistic Data Science" (OC-2018-2-23228) 2019-2022
- ENRICH4ALL (INEA/CEF/ICT/A2020/2278547) 2021-2023
- MARCELL – Multilingual Resources for CEF.AT in the legal domain" (INEA/CEF/ICT/A2017/1565710), 2018-2021
- Presidency MT Kit Automated Translation Toolkit for the EU Council Presidency (INEA/CEF/ICT/A2018/1762093) 2018-2020
- PARSEME – Parsing Multiword Expressions) –COST Action IC1207 (2013-2017)
- DRuKoLA – – Project funded by Alexander von Humbold Fundation, in partnership with IDS-Mannheim, University of București and ITI-Iași (2016-2018)
- META-NET – Multilingual Europe: a Technology Alliance (FP7, 249119 – Network of Excellence), Feb. 2010 – Jan. 2013
- METANET4U – Enhancing the European Linguistic Infrastructure (CIP-ICT-PSP.2010.6.1, 270893), Feb. 2011 – Jan. 2013
- Multilingual Web – Advancing the Multilingual Web (ICT PSP No. 250500), April 2010 – March 2012
- ACCURAT – Analysis and evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation (FP7, 248347), 2010 – 2012
- CLARIN – Common Language Resources and Technology Infrastructure, 2008 – 2011
- FlaRENet – Fostering Language Resources Network, 2008 – 2011
- COST A31 – Stability and Adaptation of Classification Systems in a Cross-Cultural Perspective, 2007 – 2010
- SEE-ERA.NET – Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages (ICT 10503 RP) 2007 – 2008

- **WISE** – An Electronic Marketplace to Support Pairs of Less Widely Studied European Languages (BSEC 009 / 05.2007) 2007 – 2008
- **Collocations en contexte** – extraction et analyse contrastive (AUF 2091RR703) 2006 – 2008
- ProLearn – PROfessional LEARNing (NoE FP6- 507310) 2004-2007
- KNOWLEDGE-WEB (NoE FP6-507482) 2004-2007
- RomNetEra – ROManian Inventory and NETworking for Integration in ERA (SSA FP6-510475) 2004-2006
- **ARS-ROCOCO** – Acquiring Reading Skills in Romanian by Comparable Corpora (bilateral project with the British Academy) 2004-2005
- **DICO-EAST** – Dictionary consultation for research and education (bilateral project with the University of Geneve) 2003-2004
- FF-POIROT – Financial Fraud Prevention-Oriented Information Resources using Ontology Technology (STREP IST-2001-38248) 2002-2005
- **EURO-CITIZEN** – Promoting European intercultural education, cross-national cooperation, and enhancing European awareness among citizens of Europe (FP5 eContent 22089) 2002-2005
- BalkaNet – Design and Development of a Multilingual Balkan WordNet (FP5, IST-2000-29388) 2001-2004
- DICOPRO – Trilateral Lexicographic Project with ISSCO – Geneva, Bulgarian Academy of Sciences; 2000-2003
- ELSNET – European Network of Excellence in Human Language Technologies (FP5, IST-1999-12127) 2000-2002
- ATAMIRI – Machine Translation Based on Trivalent Logic (Trilateral Project with Union Latine, France & University of La Paz, Bolivia) 2000-2001
- LARFLAST – Learning Foreign Language Scientific Terminology (INCO-Copernicus 977074) 1999-2001
- TELRI-II – Trans-European Language Resources Infrastructure (EC Concerted Action PL977085) 1999-2001
- **ELAN** – European Language Activity Network (MLIS Project 121) 1998-2001
- **CADBFR** – Construction Automatique de Dictionnaire Bilingue Franco-Roumain (AUPELF-UREF) 1999-2000
- AMASE – Agent based Mobile Acces for Information Services (EC ACTS AC346) 1998-1999
- CONCEDE – Consortium for Central European Dictionary Encoding (INCO-Copernicus PL961142) 1998-2000
- **FRANCIL** – Réseau FRANCophone de l'Ingénierie de la Langue (bilateral AUPELF-UREF project with LIMSI/CNRS) 1998-2000
- **HEADGEN** – Head-Driven Generator for Unification Grammars (NATO research project, with LIMSI – Paris) 1998
- **ROMANIAN INTERNET ACADEMY** – European Network of Excellence in Human Language Technologies (bilateral project with International Centre for Advanced Studies in Information Technology, Washington, SUA) 1995-1999
- ELSNET Goes EAST – European Network of Excellence in Human Language Technologies (COPERNICUS Concerted Action 200) 1995-1998
- TELRI-I – Trans-European Language Resources Infrastructure (EC Concerted Action) 1995-1997
- MULTEXT-EAST – Multilingual Text Tools and Corpora for Central and Eastern European Languages (EC COPERNICUS 106) 1995-1997
- **PAIL** – General AI Educational Environment (bilateral project with IDSIA – Lugano, Switzerland) 1993-1995

- **EGLU** – Unification-based Platform for Parsing, Transfer and Generation of Natural Language (NSF FLUENT-2 project with George Mason University-Fairfax and Massachusetts Institute of Technology-Boston, USA) 1993-1994
- **KRIL** – Knowledge-based Interlingual Generator (bilateral project with the ISSCO Institute, Geneve, Switzerland) 1993-1994
- ELSNET – European Network of Excellence in Human Language Technologies (CEC ESPRIT 6295) 1991-1995
- **RETA & ERTA** – Russian-English and English-Russian Assisted Translator (bilateral project with NPO-TPS Rusia) 1990-1992

National research projects

- ReTeRom – Resurse și tehnologii pentru dezvoltarea interfețelor om-mașina în limba română PN-III-P1-1.2-PCCDI-2018-0818, nr. 73PCCDI/2018, 2018-2021
- Robin – Roboții și Societatea: Sisteme Cognitive pentru Roboți Personali și Vehicule Autonome PN-III-P1-1.2-PCCDI-2018-0818, nr. 72PCCDI/2018, 2018-2020
- Strategia de dezvoltare a Romaniei in urmatorii 20 ani (priority program of the Romanian Academy) 2015-2017
- **COROLA** – Representative Corpus of Contemporary Romanian (priority program of the Romanian Academy) 2014-2017
- **ACCURAT-RO** – Analiza şi evaluarea corpusurilor comparabile pentru domenii ale Traducerii Automate cu mai puţine resurse (PN2, Capacităţi), coordinator: Research Institute for Artificial Intelligence, Romanian Academy, Bucharest, 2010 – 2012
- **STAR** – A System for Machine Translation for Romanian (PN2, Idei) 2009 – 2011
- **CLARIN-RO** – Infrastructură pentru resurse lingvistice interoperabile pentru limba română (PN2, Capacităţi) 2009 – 2011
- **eDTLR** – The Romanian Thesaurus Dictionary in electronic form (PN2, D91013/18.09.2007) 2007 – 2010
- **SIR-RESDEC** – Open Domain Question Answering System for Romanian and English (PN2, D11007/18.09.2007) 2007 – 2010
- **ROTEL** – Intelligent systems for the Semantic Web, based on logics, ontologies and language technologies (CEEX 29) 2005 – 2008
- **ACQ-MULT** – Acquisition and use of multilingual lexical resources (*Achiziţia şi utilizarea cunoştinţelor lexicale multilingve*) (Romanian Academy project) 2004-2005
- **IS-KS** – Strategies and Solutions for Information Society-Knowledge Society in Romania (INFOSOC) 2002
- **SI-SC** – Information Society – Knowledge Society *Societatea Informaţională – Societatea Cunoaşterii* (fundamental project of the Romanian Academy) 2001-2002
- **LINGUASTAT** – Integrated Set of Tools for Statistical Language Processing (Romanian Academy project) 1998-2001
- **WEBLEX DIC** – Generator of an XML-encoded Lexical Database from a Typographical Dictionary (Romanian Academy prioritary project) 1998-1999
- **GULIVER** – Generalized LR Parser for Unification Grammars (national research project) 1991-1993
- **PC-IURES** – Commercial version of IURES for personal computers (national research project) 1988-1991
- **PARADIGM** – Paradigmatic Morphology Learning System (national research project) 1987-1988
- **TC-LISP** – An Integrated Environment for LISP Programming: interpreter, compiler, debugger (national research project) 1984-1987

- **IURES** – System for Unrestricted Question-Answering in Natural Language (Romanian) (national research project) 1982-1983
- **SDLR** – Prototype for Natural Language (Romanian) Dialogue (national research project) 1981-1982

Activitati recente

Conferinte ievitate/ interview-uri
Data and Language-centric AI – Invited talk to International Conference on Intelligent Tutoring Systems, Bucharest, June 29-July
Dezbatere organizată de Comisia pentru știință, inovare și tehnologie a Senatului României privind progresele din domeniul inteligenței artificiale din Romania -Institutul de Cercetari pentru Iteligență Artificială "Mihai Drăgănescu"
Masa Rotunda organizata de Senatul Romaniei si Unesco – **Inteligența Artificială - etică și drepturile omului** 8.04.2024
Invited speaker la a 26ª Conferinta internationala SYNASC 2024 16-19 Septembrie 2024,Timisoara *"Ethics and Human Rights in Artificial Intelligence R&D"*

Membru în comitetele de program și reviewer pentru conferințe internationale: AIMSA 2024, CONSILR 2024
Interview Romania International -18.05.2024
Evaluator pentru European Science Foundation (proiectul 24-MUNIS-BIRD-o26)
Evaluator pentru European Science Foundation (proiectul 24-LUE-IMPACT-02)

Publicatii in2024

Anca Vasilescu, Daniela Gîfu, Dan Tufiș, Dan Cristea (eds) Linguistic Resources and Tools for Natural Language Processing. Procedings of the 18th Internatial Conference, ISSN 3045-2074, ISSN-L 3045-2074, Brașov, 11-14 December 2023

Vasile Păiș, Radu Ion, Andrei-Marius Avram, Maria Mitrofan, Dan Tufiș,
RELATE: A Modern Processing Platform for the Romanian Language; Capitol in volumul **Tools, Methods, and Solutions for the Exploration of Romanian Corpora (**Institutul European, Iași, 2024)

Unsupervised Word Sense Induction using Transformer's Attention Mechanism
Radu Ion, Vasile Păiș, Verginica Barbu Mititelu, Elena Irimia, Valentin Badea, Dan Tufiș*, submitted to Machine Intelligence and Knowledge Extraction

Underrepresented Speech Dataset from Open Data: Case Study on the Romanian Language
Vasile Păis*, Verginica Barbu Mititelu, Elena Irimia , Radu Ion and Dan Tufis,
Submitted to Applied Sciences

Vasile Păiș, Radu Ion, Elena Irimia, Verginica Barbu Mititelu, Valentin Badea, Dan Tufis,
System for the anonymization of Romanian jurisprudence. Submitted to Submitted to Applied Sciences,
Special Issue "Natural Language Processing in the Era of Artificial Intelligence"