Dr. Damir Cavar

Associate Professor
Indiana University at Bloomington

Phone: (734) 709-6039
Email: dcavar@iu.edu
Homepage: http://damir.cavar.me/
NLP Lab page: https://nlp-lab.org/
GitHub page: `https://github.com/dcavar`
ORCID: `orcid.org/0000-0002-1262-5927`

Curriculum Vitae Summary for 2024–2025

Current Positions and Roles

- **Associate Professor** at Indiana University.
- Since 2024, **Fellow of the Center for Quantum Technologies** (CQT), an NSF Industry/University Cooperative Research Center (IUCRC) including Indiana University, Purdue University, and the University of Notre Dame.
- Since 2024, **Member of the Quantum Science and Engineering Center** (QSEc) at Indiana University.
- Since 2023, **Core AI Faculty at the Luddy Artificial Intelligence Center** at Indiana University.
- Since 2018, **Fellow at the Center for Applied Cybersecurity Research (CACR)** at Indiana University.
- Since 2018, **Data Science Faculty in the Luddy School**, Computer Science Department, Indiana University.
- **Adjunct in the Russian and East-European Institute (REEI)**, School of Global & International Studies (SGIS), Indiana University.
- **Faculty Member (Voting Member)** in the Cognitive Science Program, College of Arts and Sciences (COAS), Indiana University.
- **Adjunct Associate Professor in the Department of Slavic and East European Languages and Cultures**, COAS, Indiana University
- **Advisory Board Member** of the UT Austin, Linguistic Research Center.

Most Recent Activities 2024–2025

Grants 2024–2025

– July 2024 – May 2025, Center for Quantum Technologies (CQT) grant for *Quantum-Natural Language Processing (NLP) and Machine Learning (ML)*, this is an NSF Center jointly managed by Indiana University, Purdue University, and the University of Notre Dame, full funding for one PhD-student and travel.

Related activities:

– Industry Partnership Activities (Indiana University and Cook Medical), initial meetings and preparation of joint projects.

- I received a Client-based International Projects (CLIP) Program grant for Fall 2025 to include the students of a seminar in international projects. The projects involved a cooperation with a medical team in India working on medical AI technologies, a criminology and criminal justice team in the European Union (Brussels) working on AI and NLP for forensic analysis, and a team for advertisement in Croatia working on automatic ads classification.
- Grant proposal submissions:
  - Co-PI: NSF NRT Proposal with colleagues from Physics, Mathematics, Computer Science, Cognitive Science, Chemistry, and Business: Quantum Information Science Program (PhD), submitted in Fall 2024.
  - Co-PI: Humanities and Artificial Intelligence Virtual Institute (HAVI) - Schmidt Sciences, joint proposal with a colleague from Cognitive Science and Computer Science at Indiana University at Bloomington. Submitted in Spring 2025.

Publications 2024–2025

1. Damir Cavar, Koushik Reddy Parukola (2025) *Word and Text Similarity Using Classical Word Embeddings in Quantum NLP Systems.* Satellite Workshop: Quantum Machine Learning in Signal Processing and Artificial Intelligence at the 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing. Hyderabad, India.
2. Muhammad S. Abdo, Yash Hatekar and Damir Cavar (2025) *AMWAL: Named Entity Recognition for Arabic Financial News.* In Proceedings of the FinNLP-FNP-LLMFinLegal 2025 workshop at COLING 2025.
3. Damir Cavar and Chi Zhang (2024) *Semantic Similarities using Classical Embeddings in Quantum NLP.* In Proceedings of the *IEEE Quantum Week 2024*, Montreal, Canada, September 2024.
4. Chi Zhang, Akriti Kumari, Damir Cavar (2024) *Entangled Meanings: Classification and Ambiguity Resolution in Near–Term QNLP.* In Proceedings of the *IEEE Quantum Week 2024*, Montreal, Canada, September 2024.
5. Damir Cavar, Zoran Tiganj, Ludovic Mompelat, Billy Dickson (2024) "Computing Ellipsis Constructions: Comparing Classical NLP and LLM Approaches." *Society for Computation in Linguistics* 7(1), pp. 217–226. doi: https://doi.org/10.7275/scil.2147. See (SCiL).
6. Damir Cavar, Ludovic V. Mompelat, Muhammad S. Abdo (2024) *The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications.* Pages 46–54 of Michael Hahn, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti, Saliha Muradoğlu, Rena Gao, Ryan Cotterell, Ekaterina Vylomova (eds.) Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. Association for Computational Linguistics, St Julian's, Malta. See ACL Special Interest Group on Typology (SIGTYP) 2024, collocated with the 18th Conference of the European Chapter of the Association for Computational Linguistics.

Conference Talks 2024–2025

1. Cavar, D., Koushik Reddy Parukola, Shane Sparks (2025) *Old Wine in New Bottles: Using Classical Word Embeddings in Quantum NLP Systems.* Paper to be presented at the Midwest Speech and Language Days 2025, University of Notre Dame, April 2025.

2. A. Karkala Pai, D. Cavar (2025) *A Voice-based Detection of Parkinson's Disease: Feature Selection and Classification Results.* Paper to be presented at the Midwest Speech and Language Days 2025, University of Notre Dame, April 2025.

3. R. Shrivastava, T. Sun, S.L. Anusha Chebolu, M. Kodandapani Naidu, T. Jayaprakash, J. Decatur, A. Bajpai, R. Wang, D. Cavar (2025) *Improving LLM Reasoning Through Ontology-driven Knowledge Graphs: A Comparative Study of Generating Ontologies for Medical RAGs.* Poster to be presented at the Midwest Speech and Language Days 2025, University of Notre Dame, April 2025.

4. Damir Cavar, Koushik Reddy Parukola (2025) *Word and Text Similarity Using Classical Word Embeddings in Quantum NLP Systems.* Satellite Workshop: Quantum Machine Learning in Signal Processing and Artificial Intelligence at the 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing. Hyderabad, India.

5. Entangled Meanings: Classification and Ambiguity Resolution in Near–Term QNLP. Paper presented at the Quantum AI Workshop at the IEEE Quantum Week 2024, Montreal, Canada, September 2024.

6. *Semantic Similarities using Classical Embeddings in Quantum NLP.* Poster presented at the IEEE Quantum Week 2024, Montreal, Canada, September 2024.

7. Damir Cavar, Zoran Tiganj, Ludovic Mompelat, Billy Dickson (2024) *Computing Ellipsis Constructions: Comparing Classical NLP and LLM Approaches.* Paper presented at the 2024 Meeting of the Society for Computation in Linguistics (SCiL).

8. Van Holthenrichs, Damir Cavar, Zoran Tiganj, Billy Dickson (2024) *On Ellipsis in Slavic: The Ellipsis Corpus and Natural Language Processing Results.* Paper presented at The 33rd Annual Meeting of Formal Approaches to Slavic Linguistics. Halifax, Canada.

9. *The Hoosier Ellipsis Corpus (HELC): Documenting Linguistic Dark Matter* (2024) Damir Cavar, Ludovic Mompelat, Muhammad S. Abdo. Poster presented at the Midwest Speech and Language Days at the University of Michigan in Ann Arbor, April 15-16, 2024.

10. *The Hosiers Ellipsis Corpus: Building a Corpus of Ellipsis for Arabic Natural Language Processing* (2024) Muhammad S. Abdo, Damir Cavar. Poster presented at the Midwest Speech and Language Days at the University of Michigan in Ann Arbor, April 15-16, 2024.

11. *Quantum Natural Language Processing* (QNLP) (2024) Damir Cavar, Presentation at the Quantum Day 2024 Seminar Series, organized by Quantum Technologies for Everyone (QuTE) at Indiana University Bloomington, April 14th 2024.

12. *Quantum-Natural Language Processing (NLP) and Machine Learning (ML)* (2024) Damir Cavar, Presentation at the CQT - Center for Quantum Technologies, NSF Industry/University Cooperative Research Center (IUCRC) Year 2, Phase I, Spring 2024 Industry Advisory Board Meeting, April 3-4, 2024, University of Notre Dame, South Bend, IN.

13. *Quantum Natural Language Processing and Machine Learning*. Luddy-Crane Summit on March 29, 2024 at Indiana University Bloomington.

14. *Generative AI and Knowledge Representations*. Luddy-Crane Summit on March 29, 2024 at Indiana University Bloomington.

15. *On Ellipsis in Slavic: The Ellipsis Corpus and Natural Language Processing Results*. Van Holthenrichs, Damir Cavar, Zoran Tiganj, Billy Dickson. Paper presented at The 33rd Annual Meeting of Formal Approaches to Slavic Linguistics. Halifax, Canada.

16. *The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications*. Damir Cavar, Ludovic V. Mompelat, Muhammad S. Abdo. Paper to be presented at the ACL Special Interest Group on Typology (SIGTYP) 2024, collocated with the 18th Conference of the European Chapter of the Association for Computational Linguistics, St Julian's, Malta.

17. *Ellipsis in Arabic: Using Machine Learning to Detect and Predict Elided Words*. Damir Cavar, Muhammad S. Abdo, and Billy Dickson. Paper presented at the Arabic Linguistic Society (ASAL) 37 Conference, February 2024, New York City.

18. *Building a Multilingual Ellipsis Corpus*. IU presentation, Luddy School, authors: Calvin Josenhans, John Phillips, Khai Willard, Luis Abrego, Yuchen Yang, Niko Kilo1and, Damir Cavar

Service 2024-2025

1. Member of the College Policy Committee, Spring 2024.

2. In charge of the Research Colloquium of the Department of Linguistics, Fall 2024 and Spring 2025.

3. Directing the Natural Language Processing Lab, weekly meetings during the semesters and breaks, see `https://nlp-lab.org/`.

4. Directing the Quantum NLP and AI Study Group, weekly meetings during the semesters and breaks, see `https://nlp-lab.org/quantumnlp/`.

5. Organizing the Quantum AI and NLP 2025 conference, August 2025, organizer and Conference Chair, see `https://qnlp.ai/`.

Full Curriculum Vitae

Education

Degrees

Ph.D. (*magna cum laude*), University of Potsdam, Spring 2000, Advisers: Prof. Dr. Peter Staudacher and Prof. Dr. Gisbert Fanselow

Studies

Undergraduate studies: Goethe University in Frankfurt am Main, Germany:

- Theoretical Physics and Mathematics

Graduate Studies:

Goethe University in Frankfurt am Main:

- Formal and Computational Linguistics with Prof. Dr. Helen Leuninger and Prof. Dr. Günther Grewendorf

University of Potsdam (MS and PhD studies):

- Computational Linguistics (with Prof. Dr. Gisbert Fanselow, Prof. Dr. Peter Staudacher and Prof. Dr. Jürgen Weissenborn)

Current Positions and Roles

- Since 2024, fellow of the Center for Quantum Technologies (CQT), an NSF Industry/University Cooperative Research Center (IUCRC) including Indiana University, Purdue University, and the University of Notre Dame.
- Since 2024, member of the Quantum Science and Engineering Center (QSEc) at Indiana University.
- Since 2023, Core AI Faculty at the Luddy Artificial Intelligence Center at Indiana University.
- Since 2018, fellow at the Center for Applied Cybersecurity Research (CACR) at Indiana University.
- Since 2018, Data Science Faculty in the Luddy School, Computer Science Department, Indiana University.
- Adjunct in the Russian and East-European Institute (REEI), School of Global & International Studies (SGIS), Indiana University.
- Faculty Member (Voting) in the Cognitive Science Program, College of Arts and Sciences (COAS), Indiana University.
- Adjunct Associate Professor in the Department of Slavic and East European Languages and Cultures, COAS, Indiana University
- Associate Professor (tenured), Department of Linguistics, College of Arts and Sciences, Indiana University
- Advisory Board Member of the UT Austin, Linguistic Research Center

Research Positions

Assistant Professor and **Associate Professor** at Indiana University, 2002–2006, and since 2014.

Assistant Professor and **Associate Professor** at EMU, **Director of the Institute of Language Information and Technology**, 2010–2014.

Professor, substituting, at University of Konstanz (Germany), Professor of Computational Linguistics, 2009–2010.

Assistant Professor at University of Zadar (Croatia), Professor of Computational Linguistics, 2006–2009.

Postdoc Researcher at the Berlin-Brandenburg Academy of Science, Project DWDS, Language Technology, Text-Mining, Natural Language Processing. Between 2000–2001, PI: Prof. Dr. Manfred Bierwisch and Prof. Dr. Wolfgang Klein.

Postdoc Researcher, Department of Computer Science at the Technical University in Berlin, KIT Group (Artificial Intelligence Technologies Group), working on an AI project: Verbmobil, a speech-to-speech machine translation system, my focus on dialog memory and machine learning for automatic selection of NLP-output from competing pipelines. During 2000.

Researcher (PhD student), Computer Science Institute at the University in Hamburg, Natural Language Systems Group (NATS). From 1998 to 1999. PI: Prof. Dr. Walther von Hahn and Prof. Dr. Wolfgang Menzel, Project Verbmobil, a Speech-to-Speech Machine Translation system.

Research Assistant, Berlin-Brandenburg Academy of Science, Project RULE (Rule Learning and Rule Knowledge), cognitive and computational models of communication skills in humans, animals, and insects. From 1994 to 1998. PI: Prof. Dr. Jürgen Weissenborn and Prof. Dr. Angela Friederici.

Research Assistant at the Innovationskolleg and Research Project Formal Models of Cognitive Complexity. Computational models of language processing in humans, cognitive experiments, and computational simulations. From 1994 to 1995. PI: Prof. Dr. Gisbert Fanselow, University of Potsdam.

Employment

Academic

since Summer 2023 Core AI Faculty at the Luddy Artificial Intelligence Center at Indiana University

since November 2016 Data Science Faculty in the Luddy School, Computer Science, Indiana University.

since August 2016 Associate Professor, Indiana University in the Department of Linguistics, focusing on Computational Linguistics and Natural Language Processing.

July 2014 – July 2016 Associate Research Scientist and Co-director, Indiana University in the Department of Linguistics, **Co-director** of The LINGUIST List.

2013 – 2018 Moderator of The LINGUIST List at the Institute for Language Information and Technology at Eastern Michigan University.

June 2014 Tenure and promotion to Associate Professor of Computational Linguistics at Eastern Michigan University; *excellent* for research, teaching, and service.

May 2013 – May 2014 Director of the Institute for Language Information and Technology at Eastern Michigan University.

Aug. 2011 – June 2014 Assistant Professor of Computational Linguistics (Language Technologies), Eastern Michigan University.

2006–2014 Adjunct Assistant Professor of Computational Linguistics in the Department of Linguistics at Indiana University.

2010–2011 Professor of Computational Linguistics (non-tenured substituting professor) at the University of Konstanz (Germany), temporary one-year professorship.

2009–2011 Head of the branch office of the Institute of Croatian Language and Linguistics in Zadar (headquarters in Zagreb).

2009–2010 Chair of the Linguistics Department and the Computational Linguistics Program at the University in Zadar.

2008–2011 Assistant Professor of Computational Linguistics (tenured) at the University in Zadar, Linguistics Department.

2007–2008 Visiting Associate Professor of Computational Linguistics at the University of Nova Gorica

2006–2007 Acting Chair of the English Department, University in Zadar.

2006–2007 Assistant Professor of Computational Linguistics, at the University in Zadar.

2006–2009 Head of the department for theoretical and computational linguistics at the Institute of Croatian Language and Linguistics in Zagreb.

2003–2006 Director of the Computational Linguistics Program in the Department of Linguistics at Indiana University, joint courses with Computer Science, Cognitive Science, Math.

2003–2006 Assistant Professor of Computational Linguistics (tenure track), 75% FTE in the Department of Linguistics, 25% FTE Cognitive Science Program at Indiana University.

2002–2003 Visiting Assistant Professor in the Department of Linguistics at Indiana University.

2000–2001 Postdoc Researcher, Computational Linguistics and Computational Language Resources, at the Berlin-Brandenburg Academy of Science, Project DWDS, PI: Prof. Dr. Manfred Bierwisch and Prof. Dr. Wolfgang Klein.

2000 Postdoc Researcher, Department of Computer Science at the Technical University in Berlin, KIT Group (Artificial Intelligence Technologies Group), Project Verbmobil, a speech-to-speech machine translation system.

1998–1999 Researcher (Ph.D. student), Computer Science Institute at the University in Hamburg, Natural Language Systems Group (NATS), PI: Prof. Dr. Walther von Hahn and Prof. Dr. Wolfgang Menzel, Project Verbmobil, a speech-to-speech machine translation system.

1994–1998 Researcher, Berlin-Brandenburg Academy of Science, Project RULE (Rule Learning and Rule Knowledge), Cognitive Science Project, PI: Prof. Dr. Jürgen Weissenborn and Prof. Dr. Angela Friederici.

1994–1995 Researcher at the Innovationskolleg and Research Project Formal Models of Cognitive Complexity, Cognitive Science Project, PI: Prof. Dr. Gisbert Fanselow, University of Potsdam.

Non-Academic

since 2010 Consulting in the domain of NLP and Textmining for organizations like RTI Int. and various small businesses.

2018–2023 Co-founder of Semiring Inc. (Bloomington, Indiana), an Indiana C-corp focusing on AI and NLP solutions, Semantic Web and Knowledge Graphs in various domains.

2001–2002 IT-Architect and *Project Manager*, entega Research and Development Group at the Dresdner Bank AG, Allianz Group, Frankfurt a.M., Germany. 1st level management reporting to the CTO.

1990–1994 *Freelancer, software development and network architecture*, Frankfurt a. M. Berlin (Germany).

Numerous consulting projects and software development with businesses worldwide.

Fields of Research Interest

See for more details: `https://nlp-lab.org/`

- AI, Machine Learning, Deep Learning in Natural Language Processing (NLP)
- Quantum Machine Learning and NLP, Quantum NLP
- Generative AI, Large Language Models, and Knowledge Representations with Graphs
- Knowledge Graphs and Large Language Models for Reliable Reasoning
- Computational Semantics and Pragmatics
- Natural and spoken language agents, chat-bots, dialog systems (AIs)
- Robotics with AI and NLP

Professional Activities

Reviews for Journals

- *Cognitive Science*
- *Computational Linguistics*
- *Proceedings of the National Academy of Sciences* (PNAS)
- *Lingua*

  and many others over the last 15 years

Reviews for Conferences

- *Nations of the Americas Chapter of the Association for Computational Linguistics* (NAACL)
- *FinNLP* (NLP in Financial Applications)
- *ACL* (Association for Computational Linguistics)
- *COLING* (Computational Linguistics)
- *LREC* (Language Resources and Evaluation)

- *FSMNLP* (Finite-State and NLP)
- Conference on Machine Learning, Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning (EMNLP-CoNLL)

  and numerous more over the last 15 years.

Organized Conferences and Summer Schools

Recent:

- Planning phase: Quantum AI and Natural Language Processing 2025 (QNLP) conference at Indiana University
- Midwest Speech and Language Days and the Midwest Computational Linguistics Colloquium 2016
- AARDVARC Symposium at the LSA Annual Meeting 2015 in Portland, Oregon
- Lexical Functional Grammar (LFG) 2014 at the University of Michigan
- European Lisp Symposium (ELS) 2012 at the University of Zadar
- Text Encoding Initiative (TEI) 2010 at the University of Zadar
- Summer School in Computational Linguistics 2010 at the University of Zadar
- numerous workshops and conferences at the local level at Indiana University

University Services

- Service:
  IU Bloomington Faculty Council Member (elected position) 2022-2023
  College Policy Committee Member (elected position) 2022-2023
  Various departmental services (e.g., admission committee, colloquium)

  Founder and Chair of the Linguistics Department and the Computational Linguistics Program (University of Zadar)
- Director of the Computational Linguistics Program at Indiana University (joint courses with Computer Science, Linguistics, Cognitive Science, Math); Curriculum development, 2002–2006
- Director of the LINGUIST List 2013–2018
- Director of the Institute for Language Information and Technology (ILIT)
- Member of numerous committees at the different Universities:
  - Member of the Senate at the University of Zadar (short term)
  - Member of the Octagon Committee at COAS, Indiana University, for the development of a strategic plan for the next decade
  - Admission committees, Curriculum committees at Indiana University, the University of Zadar, the University of Nova Gorica, Eastern Michigan University

Affiliations and Membership in Professional Organizations

- **Senior Member:** Association for Computing Machinery (ACM) - Lifetime Member

  - Member of the Special Interest Group for Artificial Intelligence (SIGAI)
- Member of the Institute of Electrical and Electronics Engineers (IEEE)
- Member of the Association for Computational Linguistics (ACL) over 14 years membership
- Linguistic Society of America (LSA), Lifetime Member

  Slavic Linguistic Society - Lifetime Member

Research Services

- Services for National and International Research and Funding Agencies:
  - National Science Foundation: Reviews 2008, 2009, 2011, Panel member 2006, 2014.
  - National Science Foundation in Croatia, NZZ
  - Ministry of Science of the Republic of Croatia (MZOS) 2008
  - Ministry of Science of the Republic of Slovenia 2012

Honors, Awards, & Fellowships

- Trustees Teaching Award, Indiana University, May 2004: This award is granted to one faculty member each year who has exhibited excellence in teaching at the undergraduate and/or graduate level and has served as a strong role model for graduate students.
- College of Arts and Science Summer Faculty Fellowship. Awarded amount $ 8,000
- Grant from the Faculty Research Support Program in the March 2005 competition at Indiana University for the project *A Quantitative Model of Contact-induced Language Change*
- Grant of the Republic of Slovenia, Visiting Professor at the University of Nova Gorica, 2007–2008

Miscellaneous

Technologies and Skills

Programming languages: Significant teaching experience with Python, Java, C++, Scheme, Racket, Lisp
Strong interest in: Qiskit, Pennylane, Rust

- Qiskit
- Pennylane
- Python, expert knowledge and strong teaching experience
- Rust
- Go
- Java,
- C++ and C
- Wolfram Language
- Unix-based scripting

- Motorola 68k Assembler since 1988
- Scheme and later Racket

Paradigms:

- Functional programming since my exposure to Lisp and Scheme
- Object Oriented Programming since my exposure to C++ and Java
- Quantum Algorithms since 2018

Many resources are available online and in my personal Bitbucket and GitHub repositories.

- GitHub: `https://github.com/dcavar`
- Bitbucket: `https://bitbucket.org/dcavar/`
- Personal online resources: `http://damir.cavar.me/`

Languages

- German: native
- English: native-like
- Croatian: native-like
- Polish: native-like
- Japanese: basic (University course level, minor MA topic)
- Spanish: basic (University course level)
- Azerbaijani: basic (University course level)

Research

Grants

Submitted and received in 2024:

- **Quantum Natural Language Processing** (Q-NLP), funded by the Center for Quantum Technologies (CQT) (an NSF center), Purdue University. Hired one Research Assistant to work on Quantum Algorithms for NLP and Machine Learning.
- Fellow of the Client-based International Projects (CLIP) Program through IU Global. **Applied AI technologies with an international partner organization** (government, business).
- Co-PI NSF NRT grant on Quantum Technologies in cooperation with Intelligent Systems Engineering, Physics, Chemistry, Computer Science at IU, (submitted in Fall 2024, not yet reviewed)

Submitted in 2023:

- Numerous grant proposals submitted and in preparation:
- Co-PI (senior personnel) on an NSF NRT proposal for Quantum Information Science and Engineering (QISE) at Indiana University. **Includes Computer Science, Physics, Chemistry, interdisciplinary proposal**.
- IARPA REASON, two proposals in two different teams submitted that include Accenture Federal, Michigan Tech Institute, New Jersey Institute of Technology, still waiting for results
- IARPA BENGAL, in preparation, pre-solicitation, and teaming phase, proposals related to Large Language Models in the Intelligence Community
- Industry cooperation arrangements with potential funding opportunities for research and curriculum development.

Overview of some granted funding where I am/was PI or Co-PI:

- New grant on hate-speech and anti-Semitism in social media as a multi-party funded grant to create corpora and gold-standard data sets for hate-speech or content in texts and visual information in social media content, train Deep Learning Models for the automatic detection and classification of the content. This project includes high-school and undergraduate students to educate them on how to identify and classify hate speech, how to create data resources (in a Datathon), and how to engineer and train modern AI technologies on these data sets (in a Hackathon). Many parties support this project, including private and individual donors. Also submitted to the Association for Computing Machinery (ACM), Special Interest Group AI (SIGAI) for support, as well as to Facebook Research.
- Collaborative Research Grant with Prof. Matt Josefy (Kelley School of Business, Indiana University), funded in 2014, on NLP technologies for the analysis of SEC-business reports, mapping to knowledge graphs, and development of the Deep NLP pipeline. Mining of SEC reports, network mapping of people and firms, and risk management analyses. Funded by the Office of the Office of the Vice Provost for Research at Indiana University, $ 10,000.
- Joint fellowship with Prof. Matt Josefy, Kelley School of Business at Indiana University: Fellows of the National Center for the Middle Market at the Fisher College of Business, The Ohio State University. The project is on the extraction of business data from SEC reports of Middle Market firms, developing technologies to generate networks and graph representations of middle market firm relations, people, etc. Funded with $ 30,000 in 2014.

- Conference grant for the Lexical Functional Grammar Conference 2014 in Michigan, sponsors Linguistics Program at the NSF, the Language Documentation and Endangered Languages section of the NSF. Amount: $ 12,000.
- MultiTree: Completing the Library of Language Relationships Award Number: 1227106; Principal Investigator: Damir Cavar, former PIs: Helen Aristar-Dry, Anthony Aristar. Organization: Eastern Michigan University; NSF Organization: BCS Award Date: 07/12/2012; Award Amount: $ 153,885. Project page: `http://multitree.org`. Transferred to Indiana University in 2014.
- Automatically Annotated Repository of Digital Video and Audio Resources Community (AARDVARC) Award Number: 1244713; Principal Investigator: Damir Cavar; former PI: Helen Aristar-Dry and former Co-PIs: Anthony Aristar, Damir Cavar; Organization: Eastern Michigan University; NSF Organization: BCS Award Date: 09/15/2012; Award Amount: $ 84,982. Collaborative grant: PI Douglas H. Whalen, CUNY. Project page: `http://linguistlist.org/aardvarc`, transferred to Indiana University in 2014.
- Project Manager of the text classification, text similarity, and cross-linguistic summarization sub-project of: *ATLAS: Applied Technology for Language-Aided CMS* (Role: PI of WP3, Co-PI, Funds: 3.9 mil. € gross, WP3 funds ca. 390,000 € gross; Funding period: 2010–2012. Funding agency: European Commission, a multinational project within the *Information and Communication Technologies Policy Support Programme*). Online: `http://www.atlasproject.eu/`
- Principle Investigator: *Semantic Nets and computational lexicology* (Role: PI, Funds: 75,000 €, Funding period: 2007–2011, Funding agency: Ministry for Research, Education and Sports of the Republic of Croatia).
- Co-PI: *Perception and articulation in Croatian*, funded by the Croatian Ministry of Science, Education and Sport. Project no 269-2120920-0896. Period of funding: 2007–2011. As a part of the research program: *Hrvatska Jezična Mrežna Riznica*. Primary Investigator: Małgorzata Cavar, Co-PI: Damir Cavar, partners: Antonio Oštarić (University of Zadar) and Silke Hamann (Heinrich Heine Universität Düsseldorf, Germany).
- Co-PI: *The Croatian Language Corpus: Multi-tier annotated corpus for the quantitative and qualitative analysis of language change* (Role: Co-PI, Funding period: 2005–2006, Funding agency: Ministry for Research, Education and Sports of the Republic of Croatia).
- Co-PI: NSF grant with Doug Parks (Indiana University) on developing Speech and Language Technologies for low-resourced Native American Languages.

Research Projects

This summarizes just a few of my research activities as a research scientist and Ph.D. student over the last two decades.

Verbmobil 1998–2000

Verbmobil was the largest AI and NLP project in Germany during the 90-ies, including research partners in the US, Japan, and other European countries. The goal was to develop a speech-to-speech translation system that would allow conversation partners to communicate (hear and speak) their native language while talking to somebody over the telephone. The languages used to model the systems were German, English, and Japanese.
My initial role in Computer Science (the Natural Language Group) at the University of Hamburg in the

Verbmobil project was to integrate numerous modules that a large number of research groups produced and delivered, for example speech recognition, NLP, machine translation, dialog management, knowledge representation, text to speech, on a cluster of Sun workstations using a distributed environment. My task was to compile, install, and run these modules, guaranteeing the most efficient distributed configuration for the best performance. I evaluated the single modules and the entire system end-to-end and reported to the research groups and the Ministry of Science.
In 1999, I moved from Hamburg to Berlin to the Computer Science Department at the Technical University and then to the AI group to continue my work on the project developing the dialog memory component that is necessary for anaphora resolution and other reference problems during the conversation. I also worked on a module that is based on machine learning algorithms to select the best end-to-end translation from the output of multiple parallel machine translation modules.
Spin-offs of this project were, for example, all the speech recognition developments at Carnegie Mellon, where some Verbmobil researchers ended up working. The Google Translate project was triggered by Franz Och, a former student who was working on the project at the University of Aachen. My colleague at Indiana University, Sandra Kuebler, was working with me on this project, as did numerous other colleagues who ended up at UT Austin, the University of Michigan, and so on.
Currently, the best realization of this project can be found in the new implementations of Skype by Microsoft, with plugins for simultaneous translation during a conversation or plugins for PowerPoint for instant subtitling of slide presentations into some foreign languages.
References:

Verbmobil homepage at the German Institute for Artificial Intelligence:
`http://verbmobil.dfki.de/`

DWDS 2000

At the Berlin-Brandenburg Academy of Science, my research goals in 2000 as a computational linguist were to develop XML-based technologies for the digitization of all available German publications between 1900 and 2000. The project goal was to create a text corpus with meta-annotation and linguistic analysis (part-of-speech tagged, parsed, semantically annotated) to be able to generate a 100-year overview of the German written language and to generate a digital dictionary with detailed annotations and links to reference texts.
I developed and implemented various systems for XML and XSLT-based corpus processing and annotation using the Text-Encoding Initiative XML standard. My conversion and annotation tools were implemented in C++ and were used in the DWDS project even some years ago.
For a number of years, I was active in the TEI XML community. I organized one annual summit and participated in multiple meetings.
References:

DWDS Project:
`http://www.bbaw.de/forschung/dwds/uebersicht`

TEI XML:
`http://www.tei-c.org/index.xml`

KnowledgeNet 2000–2001

In 2000, I accepted an industry-based research position to work on AI and NLP technologies for banking and insurance at the Dresdner Bank AG in Frankfurt, a branch of the Allianz Insurance Group headquartered in Munich, Germany. Besides a smaller project involving animated avatars with dialog capabilities

for customer interfaces, my main project was to develop a Knowledge Net system that maps the content of the corporate intranet texts (distributed over the globe, multi-lingual) to a graph-based search engine.
At this time, we used IBM DB2 as the SQL-based database. There were no graph DBs available for development or research purposes. We used NLP components to parse the content of texts and files, extract triple relations of the kind *subject – predicate – object*, and mapped those into handcrafted graph databases on top of IBM DB2. The texts were translated as much as possible. We had access to some machine translation technologies that emerged from cooperation with IBM on the Verbmobil project, discussed above.
The resulting system was one of the first large-scale Information Extraction systems with multi-lingual capabilities that used network and graph representations to encode the extracted information and knowledge.
During this time, I participated as an external partner on a DAML and OIL research project with the University of Karlsruhe. I got in touch with the XML technologies that later resulted in OWL, as well as various reasoning and ontology modeling approaches.
The resulting technologies from this project were handed over to a spin-off company that tried to develop and market a product, a variant of a semantic search system based on NLP, machine translation, and knowledge graphs.
I left this position to join Indiana University and launch the Computational Linguistics Program in COAS in 2002.
The work and research on this project continue today. Open Information Extraction using Deep NLP and Deep Learning, as well as advanced Graph DB representations, the Web-Ontology-Language (OWL), and reasoning, is part of my ongoing activities and grant proposals.

Croatian Language Corpus 2006–2009

In 2006, my wife was required to leave the USA due to her J1 visa status and a two-year home residency requirement. I decided to quit my position and join her in Europe. There, I was offered the position of director of computational linguistics at the Institute of Croatian Language and Linguistics in Zagreb, Croatia. I also started as an Assistant Professor at the University of Zadar, teaching computational linguistics and launching a new EU-compatible program and curriculum, as well as a brand new department with a focus on speech and language technologies.
In this context, together with colleagues at the Institute, we applied for funding for multiple research projects related to digital language resources and technologies (NLP). The initial goal was to develop the first linguistically annotated text corpus to serve data-driven methods for generating machine learning-based NLP technologies.
I used the TEI XML standard to create a large text corpus for the Croatian standard language. The corpus was used to develop dictionaries and other resources relevant to language standardization in post-war Croatia on its way to joining the EU. Language standardization was a requirement to join the EU. It was also necessary to provide translation memory for the EU and other international institutions, for example, by placing name lists and terminology standards for different domains such as law, business, and medical services. The technologies and resources developed as part of this project served all these language-related projects, and I participated in many research groups during my time there.
References:

Riznica Croatian Language Corpus:
`http://riznica.ihjj.hr/index.en.html`

TEI XML:
`http://www.tei-c.org/index.xml`

multiple related publications

GOLD 2006–2010

My colleagues from the Institute of Language Information and Technologies in Michigan invited me to participate in the General Ontology of Linguistic Description project. This was an attempt to create a Semantic Web resource (RDF) that can be used to standardize and normalize linguistic annotations and labeling in industry projects that are based on NLP. The GOLD resources have been integrated into ISOCat, which was an attempt to create an ISO standard of category labels and terms at a much larger scale.
I maintain the website and repository. The GOLD resources have been used in some of my technologies, e.g., in the morphological analyzer for Croatian, which is based on a Finite-State Transducer architecture using Semantic Web labels as output that point directly to the entries in GOLD.
References:

GOLD:
`http://linguistics-ontology.org/`

ISOCat:
`https://terms.tdwg.org/wiki/ISOcat`

multiple related publications

SNLTK 2007–2010

The Scheme Natural Language Toolkit (SNLTK) project emerged as a joint activity with my students and colleagues from Cognitive Science and Informatics to develop a library of NLP functions in Scheme and later Racket. This was presented at the European Lisp conference, and the code is available online in my GitHub code repository. I own the domain *snltk.org*, but I discontinued the web-hosting of the documentation and resources.
References:

GitHub:
`https://github.com/dcavar`

see publications

ATLAS 2008–2010

The ATLAS (Applied Technology for Language-Aided CMS) project is funded by the European Commission under the CIP ICT Policy Support Programme. Its main purpose is to facilitate multilingual web content development and management, particularly the authoring, versioning, and maintenance of multilingual Websites.
The ATLAS consortium was formed on the basis of a strong, collaborative working relationship between the partnering organizations. Tetracom IS, Atlantis Consulting, and the Institute of Technology and Development share years of experience related to supporting the development of innovative ideas in the ICT field. Furthermore, other members of the consortium from Bulgaria, Romania, Croatia, Poland, and Germany have worked together on the LT4eL and CLARIN initiatives – DFKI, University of Hamburg, Bulgarian Academy of Sciences, Institute of Computer Science of the Polish Academy of Sciences, Alexandru Ioan Cuza University of Iasi. In addition, DFKI and the University of Hamburg have collaborated for more than thirty years in various fields of research.

My task was to develop automatic document clustering and classification algorithms for the applications, as well as all the Croatian NLP components.
References:

Project homepage:
`http://www.atlasproject.eu/atlas/project/en/index.html`

Project consortium:
`http://www.atlasproject.eu/atlas/project/people/en/people.html`

see publications

GeoLing 2014–2016

GeoLing started as a web application for The LINGUIST List. I used the announcements from the mailing lists to automatically generate geo-location information for linguistic events like job announcements, conferences, institutions, and other events. The application serves as an information site that displays a world map with all the identified events and entities displayed on the map.
At the same time, this application now aggregated multiple years of data on professional mailing lists. It serves as a data repository for content of various types, meta-annotated for geo-location, and a concrete time-stamp, as well as content and type. This allows us to study the dynamic developments of one specific research discipline over time and geo-space. The extension of this system can be applied to other research disciplines and domains.
References:

Project homepage:
`http://geoling.linguistlist.org/`

AARDVARC 2010–2014

AARDVARC: Automatically Annotated Repository of Digital Audio and Video Resources Community was an NSF-funded research project that addressed the problem of not transcribing, and therefore unavailable, documentation of understudied languages by building an interdisciplinary community of linguists, anthropologists, and computer scientists to share knowledge and collaborate on the specification of a repository and suite of tools to facilitate automatic or semi-automatic transcription and analysis of audio and visual information. It provided for two workshops and a symposium to design a "take one leave one" repository and to explore recent advances in speech and video processing that allow anthropologists and linguists to break the 'transcription bottleneck' for language and cultural data. The focus on lesser-studied languages presented a new challenge for computer scientists seeking to move beyond the tools available to well-studied languages. It was an NSF-sponsored project, award number 1244713.
In the context of this project, I developed resources for low-resourced languages and experimented with speech and language technologies that could be bootstrapped from very few resources. In the domain of speech tools, I created forced-aligners for numerous languages, as well as unique resources and corpora, e.g., a first speech corpus for Eastern Chatino or a speech corpus from the AHEYM resources provided by Dov-Ber Kerler from Indiana University. Forced aligners facilitate the time alignment of transcribed speech recordings and reduce the time needed to generate large speech corpora for Automatic Speech Recognition (ASR) training.
The GORILLA project resulted from this research activity.
References:

Project homepage:
`http://info.linguistlist.org/aardvarc/`

GORILLA project:
`https://gorilla.linguistlist.org/`

see publications

GORILLA 2014–2019

The Global Open Resources and Information for Language and Linguistic Analysis (GORILLA) project resulted from the NSF-funded AARDVARC project, as discussed above.
GORILLA provides an archive, repository, and assembly line for language documentation data, corpora, computational linguistics, and speech and language resources.
The project brings together an interdisciplinary community of linguists, anthropologists, and computer scientists to collaborate on creating the tools for automatic or semi-automatic transcription and analysis of audio and visual information.
References:

Project homepage:
`http://info.linguistlist.org/aardvarc/`

GORILLA project:
`https://gorilla.linguistlist.org/`

see publications

MultiTree 2010–2014

MultiTree is a project that converts theoretical assumptions from research literature about the relationship between languages into a standardized digital format using databases and an XML tree annotation standard. The modeled theoretic assumptions are visualized using advanced visualization technologies that we developed, basing them on the JavaScript D3 library and various Python-based approaches.
MultiTree provides a unique approach to historical linguistic research, representing the most complete collection of language relationship hypotheses in a user-friendly, visually appealing, and interactive format. Not only is it fun and informative, but it is a useful resource that gathers scholarly work and makes it accessible to academics and the public alike.
MultiTree is also an innovative tool for typological analysis, especially among lesser-known languages. It facilitates interdisciplinary collaboration with linguists to reach more accurate conclusions about human language, culture, and history.
The trees in MultiTree are intended to be faithful representations of their sources, but sometimes it can be difficult to capture a scholar's intentions in a graphical representation. Whenever possible, editors have added comments to disambiguate or clarify their interpretations. However, it is always recommended that users refer to the original source for a better understanding of the scholar's hypothesis.
MultiTree aims to collect as many hypotheses about language relationships as possible so that users may compare them. The inclusion of a tree does not indicate the validity of the scholar's hypothesis or acceptance by the academic community.
Regarding contact languages (creoles, pidgins, mixed languages) and language isolates Although isolates have no known genetic affiliation, and the origins of contact languages are heavily contested, they have been included in the MultiTree database to make information about them available to scholars and to rep-

resent whatever hypothesis the original scholar is making accurately. "Trees" that include these languages do not reflect genetic affiliation unless this was the author's intention.
The initial implementation uses a hyperbolic tree visualizer based on a proprietary Java library. I converted the system to use PostgreSQL, a Python and Django-based web framework, and for visualization, a D3-based infrastructure.
The MultiTree project has been funded by the National Science Foundation (NSF grant no. 1227106).
References:

MultiTree: `http://new.multitree.org/`

see publications

LL-Map 2010–2016

LL-MAP (Language and Location – A Map Annotation Project) is a project designed to integrate language information with data from the physical and social sciences by means of a Geographical Information System (GIS). It uses advanced visualization and GIS technologies to display language information on the global map that has been digitized from existing literature describing and discussing the languages, modern and historical facts, and geo-coordinates.
The most important part of the project is a language subsystem, which relates geographical information on the area in which a language is or has been spoken to data on resources relevant to the language. Through a link to the MultiTree project, information on all proposed genetic relationships of the languages is made available and viewable in a geographic context. The system also includes ancillary information on topography, political boundaries, demographics, climate, vegetation, and wildlife, thus providing a basis upon which to build hypotheses about language movement across the territory. Some cultural information, e.g., on religion, ethnicity, and economics, is also included.
The LL-MAP system encourages collaboration between linguists, historians, archaeologists, ethnographers, and geneticists as they explore the relationship between language and cultural adaptation and change. We hope it will elicit new insights and hypotheses and that it will also serve as an educational resource. As a GIS, LL-MAP has the potential to be a captivating instructional tool, presenting complex data in a way accessible to all educational levels. Finally, as a free service available online, LL-MAP increases public knowledge of lesser-known languages and cultures, underlining the importance of language and linguistic diversity to cultural understanding and scientific inquiry.
LL-MAP started as a joint project of Eastern Michigan University (EMU) and Stockholm University in collaboration with several projects and archives in the USA, Europe, and Australia. Collaborators include PARADISEC, The Alaska Native Language Center, The Tibetan-Himalayan Digital Library, and The WALS Project, as well as noted documentary linguists. Technical development is directed by The Institute for Geospatial Research and Education (IGRE) at EMU. The project was funded by a three-year grant from the National Science Foundation.
The LL-MAP project is currently hosted and developed at Indiana University in the Department of Linguistics at The LINGUIST List.
I developed a free and open web environment from scratch after the initial funding period. The data has been converted to standards that are free and open, increasing the sustainability of the data set.
References:

LL-Map: `http://llmap.org/`

MultiTree: `http://new.multitree.org/`

see publications

FLE 2013–2019

The Free Linguistic Environment (FLE) is a project to develop a grammar engineering platform for the Lexical Functional Grammar (LFG) and related frameworks.
The project infrastructure consists of a mailing list hosted at the LINGUIST List listserv, a Bitbucket repository for research and development, and a public Bitbucket repository for the releases.
The environment is partially compatible with Xerox's Linguistic Environment. Our implementation uses Weighted Finite State Transducers in the backend and implements a unique extension of the LFG formalism that allows for probabilistic and machine learning-based grammar training or parse evaluations.
I have implemented the entire system in C++(11) using open and free standards and libraries to be fully platform-independent.
The development has been supported by a collaborative research grant from OVPR at Indiana University.
The technology is used in a joint Information Extraction project with a colleague from the Kelley School of Business. In this project, we parse and extract information from SEC business reports to map them on Knowledge Graphs and Network representations. In addition to this, the parsers and technologies are used for mapping of legal texts (from the Free Law Project, case law, and opinions) to Knowledge Graphs to enable advanced semantic search.
References:

Project page: `https://gorilla.linguistlist.org/fle/`

see publications

High-Performance Natural Language Processing

I started regular meetings at Indiana University with a focus on High-Performance NLP. The research goal of these meetings is to find ways to optimize common NLP algorithms for HPC platforms and GPU-based environments. The reading groups since the summer of 2017 on Linear Algebra and reformulating NLP solutions in terms of linear algebra implementations
References:

Project page: `http://hpnlp.org/`

see publications

OpenIE 2014–2019

Over the last few years, my research has focused on advanced Deep Linguistic technologies for use in Open Information Extraction systems. Open here refers to domain-independent. That is, the technologies are generic and only language-specific, capable of mapping semantic relations from natural language text or spoken utterances to knowledge graphs and advanced semantic representations. These representations enable semantic search, induction, reasoning, and automatic extensions of knowledge representations.
The core technologies are related to the Free Linguistic Environment mentioned above, and they include graph-based knowledge representation technologies mentioned in the Knowledge Net project above.
Funded research projects, for example, the joint cooperative project with a colleague from the Kelley School of Business, are related to the OpenIE activities, as well as joint work with students and colleagues from the Law School at Indiana University. This research is also related to the grant proposal that I am working on in the summer of 2018.
I used these technologies in seminars over the last two years. In one, we used OpenIE to extend a graph-based knowledge representation and connect it to an Alexa (Amazon AI) based interface for spoken language queries with advanced semantic processing.

During the summer of 2018, we are working with approx. 10 graduate students on Deep Learning technologies to utilize graph-based knowledge representations for applications that require an automatic generation of ontologies and taxonomies for specific domains. In addition, our focus now is to utilize these technologies, the underlying NLP and Deep Learning methods together with the graph-based knowledge representations for applications of:

- Event detection from unstructured text news (e.g. civil unrest, natural disasters, outbreaks of diseases)
- Entity and entity relation detection (mapping people, institutions, locations, and concepts onto graphs)
- Timeline detection for events
- Causal relations between events
- Validation of facts (e.g. for fake news detection, deception detection)
- Deep semantic representations of concept and concept relations

Publications

2025

Muhammad S. Abdo, Yash Hatekar and Damir Cavar (2025) AMWAL: Named Entity Recognition for Arabic Financial News. In Proceedings of the FinNLP-FNP-LLMFinLegal 2025 workshop at COLING 2025.

Damir Cavar, Koushik Reddy Parukola (2025) Word and Text Similarity Using Classical Word Embeddings in Quantum NLP Systems. Satellite Workshop: Quantum Machine Learning in Signal Processing and Artificial Intelligence at the 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing. Hyderabad, India.

2024

Two journal papers in the peer-review phase were submitted in the summer of 2024, focusing on NLP and AI technologies for Arabic and Russian.
Two conference papers on Quantum Algorithms and AI submitted to two conferences are still under review mid of December 2024.

Damir Cavar and Chi Zhang (2024) Semantic Similarities using Classical Embeddings in Quantum NLP. In Proceedings of the *IEEE Quantum Week 2024*, Montreal, Canada, September 2024. (paper, poster)

Chi Zhang, Akriti Kumari, Damir Cavar (2024) Entangled Meanings: Classification and Ambiguity Resolution in Near–Term QNLP. In Proceedings of the *IEEE Quantum Week 2024*, Montreal, Canada, September 2024. (full paper, short paper, poster)

Damir Cavar, Zoran Tiganj, Ludovic Mompelat, Billy Dickson (2024) "Computing Ellipsis Constructions: Comparing Classical NLP and LLM Approaches." *Society for Computation in Linguistics* 7(1), pp. 217–226. doi: https://doi.org/10.7275/scil.2147. See (SCiL).

Damir Cavar, Ludovic V. Mompelat, Muhammad S. Abdo (2024) The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications. Pages 46-54 of Michael Hahn, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti, Saliha Muradoğlu, Rena Gao, Ryan Cotterell, Ekaterina Vylomova (eds.) Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP. Association for Computational Linguistics, St Julian's, Malta. See ACL Special Interest Group on Typology (SIGTYP) 2024, colocated with the 18th Conference of the European Chapter of the Association for Computational Linguistics. (full paper)

2022

Damir Cavar, Ali Aljubailan, Ludovic Mompelat, Yuna Won, Billy Dickson, Matthew Fort, Andrew Davis and Soyoung Kim (2022) *Event Sequencing Annotation with TIE-ML* In proceedings of The Eighteenth Joint ACL – ISO Workshop on Interoperable Semantic Annotation (ISA-18 2022), at LREC 2022 in Marseille, France.

Günther Jikeli, Damir Cavar, Weejeong Jeong, Daniel Miehling, Pauravi Wagh, Denizhan Pak (2022) *Toward an AI Definition of Antisemitism?* Pages 193-212 in M. Hübscher and S. von Mering (eds.) Antisemitism on Social Media. Routledge, New York.

Ali Abdulaziz Aljubailan, Damir Cavar (2022) *Time, Language, Knowledge, and Knowledge Management in the Era of Big Data.* To appear in proceedings of the 4th Biennial U-M International Conference on Arabic Applied Linguistics. University of Michigan, Ann Arbor.

2021

Damir Cavar, Billy Dickson, Ali Aljubailan, Soyoung Kim (2021) "Temporal Information and Event Markup Language: TIE-ML Markup Process and Schema Version 1.0," In Proceedings of SEMAPRO 2021, Barcelona, Spain.

2020

Gunther Jikeli, Damir Cavar, Weejoeng Joeng, Daniel Miehling, Denishan Pak, Pauravi Wagh (2020) *Towards an AI Definition of Antisemitism?* Manuscript, Indiana University.

2019

Damir Cavar et al. (2019) Bootstrapping Legal Information Extraction and Knowledge Graphs. Paper submitted to a major conference in the domain of Computational Linguistics and AI, summer 2019.

Damir Cavar et al. (2019) Generating Dynamic Knowledge Graphs from Text using Deep and Broad NLP. Paper submitted to a major conference in the domain of Computational Linguistics and AI, summer 2019.

Damir Cavar et al. (2019) Geometry of Cross-Linguistic Lexical Similarities using Pre-computed Word Vector Models. Paper submitted to a major conference in the domain of Computational Linguistics and AI, summer 2019.

Damir Cavar et al. (2019) Standardized Parallel NLP Pipeline Microservice Architecture. Paper submitted to a major conference in the domain of Computational Linguistics and AI, summer 2019.

Damir Cavar (2019) Measuring Lexical Semantic Variation using Word Embeddings. To appear in TBA (book chapter).

Damir Cavar, Oren Baldinger, Joshua Herring, Umang Mehta, Yiwen Zhang, Shantanu Bedekar, Shreejith Panicker (2019) An Annotation Encoding Schema for Natural Language Processing using JSON: JSON-NLP Schema Version 0.1. Technical Report, NLP Lab, Indiana University, Version 1.0 from November 2018.

2018

Damir Cavar, Joshua Herring, Anthony Meyer (2018) *Case Law Analysis using Deep NLP and Knowledge Graphs*. in Proceedings of the LREC 2018, paper presented at the 1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph (LegalKG), LREC 2018, in Miyazaki, Japan.

Damir Cavar, Matt Josefy (2018) *Mapping Deep NLP to Knowledge Graphs: An Enhanced Approach to Analyzing Corporate Filings with Regulators*. In Proceedings of The First Financial Narrative Processing Workshop (FNP 2018), LREC 2018 in Miyazaki, Japan.

2016

Damir Cavar, Lwin Moe, Hai Hu, Kenneth Steimel (2016) Preliminary Results from the Free Linguistic Environment Project. Pages 161–181 in D. Arnold, M. Butt, B. Crysmann, T. Holloway-King, S. Müller (eds.) *Proceedings of the Joint 2016 Conference on Head-driven Phrase Structure Grammar and Lexical Functional Grammar*. CSLI Publications.

Małgorzata E. Ćavar, Damir Cavar, Hilaria Cruz (2016) *Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR*. In Proceedings of the LREC 2016, Portorož, Slovenia.

Damir Cavar, Małgorzata E. Ćavar, Lwin Moe (2016) *Global Open Resources and Information for Language and Linguistic Analysis (GORILLA)*. In Proceedings of the LREC 2016, Portorož, Slovenia.

Małgorzata E. Ćavar, Damir Cavar, Dov-Ber Kerler, Anya Quilitsch (2016) *Generating a Yiddish Speech Corpus, Forced Aligner and Basic ASR System for the AHEYM Project.* In Proceedings of the LREC 2016, Portorož, Slovenia.

2014

Damir Cavar and Malgorzata Cavar (2014) Visualization of Language Relations and Families: MultiTree. In: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds.) *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14), May 26–31. Reykjavik, Iceland, European Language Resources Association (ELRA), ISBN 978-2-9517408-8-4.

2012

Anelia Belogay, Damir Cavar, Dan Cristea, Diman Karagiozov, Svetla Koeva, Roumen Nikolov, Maciej Ogrodniczuk, Adam Przepiórkowski, Polivios Raxis, and Cristina Vertan (2012) *i-Publisher, i-Librarian and EUDocLib – linguistic services for the Web*. In Piotr Pęzik (ed.) Corpus Data across Languages and Disciplines, volume 28 of Łódź Studies in Language, pages 203–212. Peter Lang

Damir Cavar, Helen Aristar-Dry, Anthony Aristar (2012) Large Mailing List Corpora: Management, Annotation and Repository. In *LREC 2012 Proceedings of the workshop on Challenges in the management of large corpora*.

Damir Cavar, Dunja Brozović Rončević (2012) *Riznica: The Croatian Language Corpus*. In *Prace Filologiczne* LXIII (63), Wydział Polonistyki Uniwersytetu Warszawskiego, Warsaw. Pages 51-66. ISSN: 0138-056.

2011

Damir Cavar, Tanja Gulan, Damir Kero, Franjo Pehar, Pavle Valerjev (2011) The Scheme Natural Language Toolkit (SNLTK): NLP libraries for R6RS and Racket. *In Proceedings of the 4th European Lisp Symposium*, Hamburg University of Technology, pp. 58–61.

Damir Cavar, Melanie Seiß (2011) Clitic Placement, Syntactic Discontinuity, and Information Structure. In: M. Butt and T. Holloway King (eds.) *Proceedings of the LFG 2011 conference* in Hong Kong, p. 131–151. Online Proceedings ISSN 1098-6782

2010

Damir Cavar (2010) On Statistical Metrics for Selection and Phrasality. In T. Hanneforth and G. Fanselow (eds.) *Language and Logos: Studies in Theoretical and Computational Linguistics*. Studia grammatica, 72, p. 393–406. Akademie Verlag, Berlin. ISBN 978-3050049311

2009

Damir Cavar, Ivo-Pavao Jazbec, Siniša Runjaić (2009) Efficient Morphological Parsing with a Weighted Finite State Transducer. *Informatica* 33/1, pp. 107–113. Website of the journal. ISSN: 0350-5596

Damir Cavar, Ivo-Pavao Jazbec, Tomislav Stojanov (2009) CroMo – Morphological Analysis for Standard Croatian and its Synchronic and Diachronic Dialects and Variants. In: J. Piskorski, B. W. Watson, A. Yli-Jyrä (eds.) *Finite-State Methods and Natural Language Processing*, Frontiers in Artificial Intelligence and Applications 19, IOS Press, pp. 183–190. ISBN 978-1-58603-975-2

Damir Cavar, Ivo Pavao Jazbec, Bruno Nahod (2009) Struktura i razvoj baze podataka za potrebe projekta Hrvatsko strukovno nazivlje (STRUNA) - projekt koordinacije. Pages 311–317 in: N. Ledinek, M. Žagar Karer and M. Humar (eds.) *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU. ISBN: 978-961-254-158-3

2008

Damir Cavar, Ivo-Pavao Jazbec, Siniša Runjaić (2008) *Interoperability and Rapid Bootstrapping of Morphological Parsing and Annotation Automata*. In Proceedings of IS-LTC 08, Ljubljana, Slovenia. Website of the proceedings

Damir Cavar, Ivo-Pavao Jazbec, Tomislav Stojanov (2008) *CroMo – Morphological Analysis for Standard Croatian and its Synchronic and Diachronic Dialects and Variants*. In Proceedings of FSMNLP 2008, Ispra, Italy. Website of the Conference

Paul Rodrigues, Damir Cavar (2008) Learning Arabic Morphology With Information Theory. *Proceedings from the Annual Meeting of the Chicago Linguistic Society* (CLS 41), Volume 41, No. 2. Chicago, IL, USA. Pages 49–60. CLS Online Journal Online. ISSN: 0577-7240

2007

Damir Cavar, Dunja Brozović Rončević (2007) *Grammaticality judgments and language usage data*. In Proceedings of the fourth Corpus Linguistics conference 2007, Corpus and Cognition Colloquium: The relation between natural and experimental language data; Birmingham, July 27–30, 2007. Just the abstract.

Paul Rodrigues, Damir Cavar (2007) "Learning Arabic Morphology Using Statistical Constraint Satisfaction Models." Pages 63–76 in: E. Benmamoun (ed.) *Perspectives on Arabic Linguistics* XIX. Current Issues in Linguistic Theory 289. John Benjamins, Amsterdam. Google Books. ISBN: 978-90-272-4804-6

2006

Damir Cavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, Giancarlo Schrementi (2006) On Unsupervised Grammar Induction from Untagged Corpora. Pages 55–71 in: P. Kaszubski (ed.) *PSiCL: Poznań Studies in Contemporary Linguistics* 41, Adam Mickiewicz University, Poznań, Poland. Online Journal. ISBN: 73-7208-165-4

Damir Cavar, Paul Rodrigues, Giancarlo Schrementi (2006) *Unsupervised Morphology Induction for Part-of-Speech Tagging*. Penn Working Papers in Linguistics. Volume 12.1. Philadelphia, PA, Seiten 29–41. Online Working Papers

2005

Damir Cavar, Paul Rodrigues, Giancarlo Schrementi (2005) *Using Morphological and Distributional Cues for Inductive Part-of-Speech Tagging*. Proceedings of the Midwest Computational Linguistics Colloquium (MCLC) 2005 at the Ohio State University, Columbus, OH. Online

2004

Damir Cavar, Paul Rodrigues, Giancarlo Schrementi (2004) *Syntactic Parsing Using Mutual Information and Relative Entropy*. Proceedings of Midwest Computational Linguistics Colloquium (MCLC) 2004. Online Site

Damir Cavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, Giancarlo Schrementi (2004) On Statistical Bootstrapping. In: W.G. Sakas (ed.) *Proceedings of the First Workshop on Psycho-computational Models of Human Language Acquisition*, held in cooperation with COLING 2004, Geneva, Pages 9–16. ACL Anthology Online

Damir Cavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, Giancarlo Schrementi (2004) Alignment Based Induction of Morphology Grammar and its Role for Bootstrapping. In: G. Jäger, P. Monachesi, G. Penn and S. Wintner (eds.) *Proceedings of Formal Grammar 2004*, Nancy, Pages 47–62. Online Proceedings

2002

Chris Wilder and Damir Cavar (2002) Verb Movement, Cliticization, and Coordination. Pages 365–375 in: P. Kosta and J. Frasek (eds.) *Current Approaches to Formal Slavic Linguistics*. Linguistik International Series Vol. 9, Peter Lang: Frankfurt a.M. Google Books. ISBN: 978-3-631-50311-9

Sebastian Brandt, Damir Cavar and Uta Störl (2002) *A Real Live Web Service using Semantic Web Technologies: Automatic Generation of Meta-Information*. In: Proceedings of "On The Move Towards Meaningful Internet Systems" (DOA, ODBASE, CoopIS'02), Irvine, California.

Damir Cavar and Richard Kauppert (2002) "Strategien für die Implementierung IT-basierter KM-Lösungen: Minimal Invasive Systeme". In: C. Prange (ed.) *Organisationales Lernen und Wissensmanagement – Fallstudien aus der Unternehmenspraxis*. Gabler Verlag. Google Books. ISBN: 3-409-11762-8

Damir Cavar and Uta Störl (2002) Automatic Generation of Meta Tags for Intra-Semantic-Web. Pages 67–77 in: R. Tolksdorf, R. Eckstein (eds.) *XML Technologien für das Semantic Web* – XSW 2002. Gesellschaft für Informatik, Berlin, Germany, Lecture Notes in Informatics, Vol. 14. Google Books. ISBN 3-88579-343-1

2001

Gisbert Fanselow and Damir Cavar (2001) Distributed Deletion. Pages 65–108 in: A. Alexiadou (ed.) *Theoretical Approaches to Universals*. Benjamins, Amsterdam. Google Books. ISBN: 90-272-2770-5

2000

Gisbert Fanselow and Damir Cavar (2000) Remarks on the economy of pronunciation. Pages 107–150 in: G. Müller and W. Sternefeld (eds.) *Competition in Syntax*, Studies in Generative Grammar 49. Google Books. ISBN 3-11-016945-2

Damir Cavar, Alexander Geyken, Gerald Neumann (2000) Digital Dictionary of the 20th Century German Language. In: T. Erjavec and J. Gros (eds.) *Jezikoslovne Tehnologije za Slovenski Jezik*. Proceedings of JS 2000, Ljubljana: Institut Jožef Stefan. Online Proceedings. ISBN 961-6303-25-2

Damir Cavar, Uwe Küssner, Dan Tidhar (2000) "From Off-line Evaluation to On-line Selection". Pages 599–612 in: W. Wahlster (ed.) *Verbmobil: Foundations of Speech to Speech Translation*, Springer Verlag. Das Verbmobil Buch. ISBN: 3-540-67783-6

Damir Cavar, Uwe Küssner, Dan Tidhar (2000) "From Human Evaluation to Automatic Selection of Good Translations". In: Proceedings of the Second International Conference on Language Resources and Evaluation, LREC 2000 – Workshop on the Evalution of Machine Translation, Athens. CiteSeerX Online

1999

Gisbert Fanselow, Matthias Schlesewsky, Damir Cavar, Reinhold Kliegl (1999) *Optimal parsing, syntactic parsing preferences, and Optimality Theory*. (Rutgers Optimality Archive Online Publication). ROA Online

Jürgen Weissenborn, Barbara Höhle, Dorothea Kiefer, Damir Cavar (1998) Children's Sensitivity to Word-Order Violations in German: Evidence for very Early Parameter-Setting. Pages 756–767 in: A. Greenhill, M. Hughes, H. Littlefield and H. Walsh (eds.) *BUCLD 22: Proceedings of the 22nd annual Boston University Conference on Language Development*, Somerville, Cascadilla Press. Online. ISBN: 978-1-57473-032-6

Damir Cavar and Chris Wilder (1999) "Clitic Third in Croatian". Pages 429–468 in: H. v. Riemsdijk (ed.) Clitics in the Languages of Europe. Mouton de Gruyter, Berlin. Google Books. ISBN: 3-11-015751-9

First publication (unabridged): H. v. Riejmsdijk and L. Hellan (1994) (eds.) *Clitics: Their Origin, Status and Position*. Eurotype Woking Papers, Theme Group 8, Vol. 6.

Damir Cavar (1999) *Aspects of the Syntax-Phonology Interface*. PhD thesis, University of Potsdam.

1998

Damir Cavar and Gisbert Fanselow (1998) "Discontinuous constituents in Slavic and Germanic languages", Mscr., University of Potsdam. Online publication.

Damir Cavar and Wolfgang Menzel (1998) VERBMOBIL: A Speech-to-Speech Translation System. Pages 25–28 in: T. Erjavec and J. Gros (eds.) Language Technologies for the Slovene Language: proceedings of the conference. Jozef Stefan Institute, Ljubljana. Online Proceedings. ISBN: 961-6303-00-7

Damir Cavar and Chris Wilder (1998) "Auxiliaries in Serbian/Croatian and English" Pages 3-12 in: U. Junghanns and G. Zybatow (eds.) *Formale Slavistik*. Vervuert: Frankfurt a.M. Online. ISBN: 3-89354-267-1

1996

Damir Cavar (1996) "On Clitics in Croatian: Syntax or Phonology?" Paper presented at the "Workshop on the Syntax, Morphology and Phonology of Clitics" In: *ZAS-Working Papers in Linguistics* 6, Papers on clitics (Oct. 1996), Pages: 51–65. Online

1994

Damir Cavar (1994) *Minimalist Aspects of the Syntax of Closed Class Elements*, Diploma thesis, University of Potsdam.

Damir Cavar and Chris Wilder (1994) "Clitic Third in Croatian" In: *Linguistics in Potsdam* No. 1, 25–63. Online

Damir Cavar and Chris Wilder (1994) "X^0-Bewegung und Ökonomie" Pages 11–32 in: B. Haftka and C. M. Schmitt (eds.) *Was determiniert Wortstellungsvariationen? Studien zu einem Interaktionsfeld von Grammatik, Pragmatik und Sprachtypologie*. Westdeutscher Verlag: Opladen. ISBN: 9783531124902

Chris Wilder and Damir Cavar (1994) "Word Order Variation, Verb Movement and Economy Principles". In: *Studia Linguistica* 48.1, pp. 46–86. DOI

Damir Cavar and Chris Wilder (1994) "Long Head Movement? Verb-Movement and Cliticization in Croatian" In: *Lingua* 93.1, pp. 1–58. DOI

1993

Chris Wilder and Damir Cavar (1993) "Word Order Variation, Verb Movement and Economy Principles". Working papers *Sprachwissenschaft in Frankfurt* 10, Frankfurt a.M.

1992

Damir Cavar and Chris Wilder (1992) "Long Head Movement? Verb-Movement and Cliticization in Croatian". Working papers *Sprachwissenschaft in Frankfurt* 7, Frankfurt a.M.

Scientific Software

Most projects are public on my GitHub repo (`https://github.com/dcavar`), including:

- Python Teaching Notebooks as Jupyter Notebooks
- The Hoosier Ellipsis Corpus with code for Large Language Models

HooSIER: HooSIER Semantic Information ExtractoR. A text to Knowledge Graph system.

JSON-NLP, a standard for outputs of Natural Language Processing pipelines.

NLP-Ensemble: an industry-standard High-Performance Computing Natural Language Processing environment based on a scalable RESTful Microservice architecture and Deep NLP and Knowledge Driven content analysis of text and visual content.

Scheme Natural Language Toolkit (`http://www.snltk.org`, see GitHub repos for dcavar), A library of natural language processing functions, quantitative text processing methods, and visualization methods written in Scheme.

Language Identification Algorithm (www.cavar.me/damir/LID), A set of algorithms to train a language identifier for text and a standard deviation-based recognition algorithm.

TextStat (www.cavar.me/damir/textstat), A library of functions implemented in Python 3 for statistical analysis of text, generation of N-gram models, visualization as dot-graphs, and more.

Charty (www.cavar.me/damir/charty), a syntactic parser using context-free grammars, a variant of an Early Parser, implemented in Python 2.x, Python 3.x and Scheme.

IPA Transcriber (ltl.emich.edu/ dcavar/phonemic), automatic transcription of language-specific text to the International Phonetic Alphabet, implemented in Scheme.

Free Linguistic Environment: a context-free grammar parser implemented as a weighted finite state transducer to enable probabilistic CFG-parsing, combined with an algorithm for Probabilistic Directed Acyclic Graphs, to enable Probabilistic or Graded Unification over feature graphs.

ELAN2split: a corpus generator that uses ELAN time-aligned speech corpus annotations and WAV audio files to generate corpora for Automatic Speech Recognition training.

many more software components and tools for linguistic studies, language processing, and data analysis in the context of research projects and teaching activities

Presentations

Invited lectures and conference presentations: most recent

2025 Cavar, D., Koushik Reddy Parukola, Shane Sparks (2025) Old Wine in New Bottles: Using Classical Word Embeddings in Quantum NLP Systems. Paper to be presented at the Midwest Speech and Language Days 2025, University of Notre Dame, April 2025.

2025 A. Karkala Pai, D. Cavar (2025) A Voice-based Detection of Parkinson's Disease: Feature Selection and Classification Results. Paper to be presented at the Midwest Speech and Language Days 2025, University of Notre Dame, April 2025.

2025 R. Shrivastava, T. Sun, S.L. Anusha Chebolu, M. Kodandapani Naidu, T. Jayaprakash, J. Decatur, A. Bajpai, R. Wang, D. Cavar (2025) Improving LLM Reasoning Through Ontology-driven Knowledge Graphs: A Comparative Study of Generating Ontologies for Medical RAGs. Poster to be presented at the Midwest Speech and Language Days 2025, University of Notre Dame, April 2025.

2025 Davis, A.S., B. Dickson, D. Cavar, D. Valdez, F.M. Tyers (2025) Advancing Adverse Drug Event Detection in Social Media Through Knowledge Graph and GraphRAG LLM Architectures. 2025 AAHB Annual Scientific Meeting, San Diego, CA. (Poster)

2024 Entangled Meanings: Classification and Ambiguity Resolution in Near–Term QNLP. Poster presented at the IEEE Quantum Week 2024, Montreal, Canada, September 2024.

2024 Entangled Meanings: Classification and Ambiguity Resolution in Near–Term QNLP. Paper presented at the Quantum AI Workshop at the IEEE Quantum Week 2024, Montreal, Canada, September 2024.

2024 Semantic Similarities using Classical Embeddings in Quantum NLP. Poster presented at the IEEE Quantum Week 2024, Montreal, Canada, September 2024.

2024 Damir Cavar, Zoran Tiganj, Ludovic Mompelat, Billy Dickson (2024) Computing Ellipsis Constructions: Comparing Classical NLP and LLM Approaches. Paper presented at the 2024 Meeting of the Society for Computation in Linguistics (SCiL).

2024 Van Holthenrichs, Damir Cavar, Zoran Tiganj, Billy Dickson (2024) On Ellipsis in Slavic: The Ellipsis Corpus and Natural Language Processing Results. Paper presented at The 33rd Annual Meeting of Formal Approaches to Slavic Linguistics. Halifax, Canada. (abstract, slides)

2024 *The Hoosier Ellipsis Corpus (HELC): Documenting Linguistic Dark Matter* (2024) Damir Cavar, Ludovic Mompelat, Muhammad S. Abdo. Poster presented at the Midwest Speech and Language Days at the University of Michigan in Ann Arbor, April 15-16, 2024.

2024 *The Hosiers Ellipsis Corpus: Building a Corpus of Ellipsis for Arabic Natural Language Processing* (2024) Muhammad S. Abdo, Damir Cavar. Poster presented at the Midwest Speech and Language Days at the University of Michigan in Ann Arbor, April 15-16, 2024.

2024 *Quantum Natural Language Processing* (QNLP) (2024) Damir Cavar, Presentation at the Quantum Day 2024 Seminar Series, organized by Quantum Technologies for Everyone (QuTE) at Indiana University Bloomington, April 14th 2024.

2024 *Quantum-Natural Language Processing (NLP) and Machine Learning (ML)* (2024) Damir Cavar, Presentation at the CQT - Center for Quantum Technologies, NSF Industry/University Cooperative Research Center (IUCRC) Year 2, Phase I, Spring 2024 Industry Advisory Board Meeting, April 3-4, 2024, University of Notre Dame, South Bend, IN.

2024 *Quantum Natural Language Processing and Machine Learning*. Luddy-Crane Summit on March 29, 2024 at Indiana University Bloomington.

2024 *Generative AI and Knowledge Representations*. Luddy-Crane Summit on March 29, 2024 at Indiana University Bloomington.

2024 *On Ellipsis in Slavic: The Ellipsis Corpus and Natural Language Processing Results*. Van Holthenrichs, Damir Cavar, Zoran Tiganj, Billy Dickson. Paper presented at The 33rd Annual Meeting of Formal Approaches to Slavic Linguistics. Halifax, Canada.

2024 *The Typology of Ellipsis: A Corpus for Linguistic Analysis and Machine Learning Applications*. Damir Cavar, Ludovic V. Mompelat, Muhammad S. Abdo. Paper to be presented at the ACL Special Interest Group on Typology (SIGTYP) 2024, colocated with the 18th Conference of the European Chapter of the Association for Computational Linguistics, St Julian's, Malta.

2024 *Ellipsis in Arabic: Using Machine Learning to Detect and Predict Elided Words*. Damir Cavar, Muhammad S. Abdo, and Billy Dickson. Paper presented at the Arabic Linguistic Society (ASAL) 37 Conference, February 2024, New York City. (slides)

2024 *Building a Multilingual Ellipsis Corpus*. IU presentation, Luddy School, authors: Calvin Josenhans, John Phillips, Khai Willard, Luis Abrego, Yuchen Yang, Niko Kilo1and, Damir Cavar

2023 *Artificial Intelligence and Entertainment Webinar*, organized by Gotham Group, Los Angeles, Hollywood. Discussing Large Language Models, AI, and Language Technologies with The Writers Union, Producers, Directors.

2019 Damir Cavar (2019) *Deep and Broad NLP for Big Data and Knowledge Graph Generation*, paper presented at the Data Science Colloquium at the School of Informatics, Computation, and Engineering at Indiana University, March 22nd, 2019.

2019 Damir Cavar (2019) Semantic Information Extraction and Generation of Dynamic Knowledge Graphs. Paper presented at the University of Illinois at Urbana-Champaign.

2018 *Deep Linguistics and Deep Learning for Natural Language Processing: A Practical AI System (Conversational Agent)*. City University of New York (CUNY) Graduate Center. March 2018.

2017 *An infrastructure for Global Open Resources and Information for Language and Linguistic Analysis* (GORILLA). University of North Texas in Denton. Feb. 9th 2017.

2010 Plenary speaker: TEI 2010 (Annual meeting of the Text Encoding Initiative), University of Zadar

2009 North-West University, CTexT, Potchefstroom, South Africa: a. *On bootstrapping of linguistic features for bootstrapping grammars*; b. *Finite State Automata and Regular Languages*; c. *Document Classification and Clustering using KNN*

2009 Invited speaker: EACL 2009 in Athens: *On bootstrapping of linguistic features for bootstrapping grammars*

2009 Invited speaker: 10^{th} Meeting in Szklarska Poreba (Poland): *On the induction of linguistic categories and learning grammars*

2007 Invited speaker: University in Ljubljana: *Dynamic Language Models*

Presentations at conferences and workshops

2023

- *Understanding Ellipsis in Language: A Comparative Analysis of SOTA NLP and Large Language Models*, Cognitive Science Lunch talk at Indiana University at Bloomington, October 2023.
- *Artificial Intelligence and Entertainment Webinar*, organized by Gotham Group, Los Angeles.
- *Distributed Deletion, Syntax, and Knowledge Representation* (2023) Paper presented at the Gisbert Fanselow's Contributions to Syntactic Theory and GGS 47, Berlin, Germany.
- *Automated Hate Speech Detection - The Importance of Precise Datasets Including a Calling-Out-Bias Label*, (2023) Poster presented at the Indiana University AI Day, Bloomington, Indiana.

2022

- On Tense interpretation in Croatian, Polish, Russian, Ukrainian — A Corpus Study and Computational Model. Slavic Linguistics Society 17, Sapporo, Japan; September 2022.
- Event Sequencing Annotation with TIE-ML. Paper presented at The Eighteenth Joint ACL - ISO Workshop on Interoperable Semantic Annotation (ISA-18 2022), Marseille, France; June 2022.

2021

- Computational Semantics and Reasoning using Knowledge Graphs and Multi-modal Information Sources for Knowledge Extraction, presented at the Computational Linguistics Seminar at the University of North Carolina at Chapel Hill, October 6th, 2021.
- Temporal Information and Event Markup Language: TIE-ML Markup Process and Schema Version 1.0, SEMAPRO 2021 in Barcelona, Spain.

2019

- Damir Cavar (2019) *On NLP and Knowledge Graphs, TBS,* Keynote speaker at the 3rd Asian Conference on Artificial Intelligence Technology (ACAIT 2019), Chongqing University of Technology (CQUT), China.
- Damir Cavar and Elain Monaghan (2019) *Mapping News Article Content to Knowledge and Event Graph Representations for Validation and Reasoning*. "Media Ethics: Human Ecology in a Connected World" 20th annual convention of the Media Ecology Association, Toronto, (26) 27–30 June.
- Damir Cavar (2019) *Deep and Broad NLP for Big Data and Knowledge Graph Generation*, paper presented at the Data Science Colloquium at the School of Informatics, Computation, and Engineering at Indiana University, March, 22nd, 2019.
- Damir Cavar (2019) Semantic Information Extraction and Generation of Dynamic Knowledge Graphs. Paper presented at the University of Illinois at Urbana-Champaign.

2018

- *Deep Linguistics in Chatbots*, ICWSM-18 Workshop on Chatbots, The 12th International AAAI Conference on Web and Social Media, Stanford, California. June 2018.

- *Knowledge Graphs and Open IE*, The Annual Conference on Data, Information, and Society, Nanjing, China, July 2018.
- *Deep Learning Models of Linguistic Similarity*, Montana State University, Billings, Montana, May 2018.
- *Deep Linguistics and Deep Learning for Natural Language Processing: A Practical AI System (Conversational Agent)*. City University of New York (CUNY) Graduate Center. March 2018.

Canceled presentations in 2018:

- *Keynote lecture at the 32nd Annual Symposium on Arabic Linguistics*. Arizona State University, Tempe Arizona.
  `https://linguistlist.org/confservices/customhome.cfm?CFID=4563ad66-7fa4-4680-bf33-f328bdfeb535&emeetingid=5902JA445856625840A050441`
- *Case Law Analysis using Deep NLP and Knowledge Graphs. Paper presented at the 1st Workshop on Language Resources and Technologies for the Legal Knowledge Graph (LegalKG)*, LREC 2018, Miyazaki, Japan.
- *Mapping Deep NLP to Knowledge Graphs: An Enhanced Approach to Analyzing Corporate Filings with Regulators*, LREC 2018, Miyazaki, Japan.

2017

- *Computational Semantics and Computational Pragmatics in NLP for Professional Domain Sublanguage: Processing Medical Language*. Paper presented at the CHSR WIP, Indiana University, School of Medicine. November 13th, 2017.
- *The Free Linguistic Environment: High-performance Deep and Broad Coverage Multi-lingual NLP*. Paper presented in the Intelligent & Interactive Systems Talk Series at the School of Informatics, Computation, and Engineering, Indiana University. April 24th, 2017.
- *The Free Linguistic Environment: Theoretical Concepts and Basic Architecture*. Paper presented at Cornell University, March 30th, 2017
- *Speech Corpora and Technologies for Yiddish (on the AHYEM Project)*. Joint work with Malgorzata E. Cavar, Dov-Ber Kerler, and Anya Quilitzsch. Paper presented at the Linguistics Research Center, the Texas Language Center, and the Department of Germanic Studies at the University of Texas at Austin. Feb. 23rd, 2017.
- *An infrastructure for Global Open Resources and Information for Language and Linguistic Analysis* (GORILLA). Paper presented at the University of North Texas in Denton. Feb. 9th 2017.
- *On the Free Linguistic Environment: Parsing and Corpus Annotation*, together with Lwin Moe, Hai Hu. Paper presented at the 15th International Workshop on Treebanks and Linguistic Theories, Bloomington, Indiana.

2016

On Recursion. Cognitive Science Colloquium, Indiana University, Nov. 16th, 2016, Bloomington, Indiana.

On Split Islands. 11th Slavic Linguistic Society Annual Meeting. Sept. 24th, 2016. Toronto, Canada.

Preliminary Results from the Free Linguistic Environment Project. Paper presented at the HeadLex16, 24.–29. July 2016, Warsaw, Poland.

Plenary panel that sets out needs/issues from four perspectives. Tools & Methods Summit, University of Melbourne, Australia, 1st–3rd of June, 2016.

Generating speech tools and corpora for unwritten endangered languages: Chatino. Paper presented at the Tools & Methods Summit, University of Melbourne, Australia, 1st–3rd of June, 2016.

Bridging between Documentary Linguistics, Computational Linguistics, Theoretical Linguistics, and NLP. Paper presented at the Tools & Methods Summit, University of Melbourne, Australia, 1st–3rd of June, 2016.

Concept and work behind GORILLA. Paper presented at the Tools & Methods Summit, University of Melbourne, Australia, 1st–3rd of June, 2016.

On the Role of Linked Open Language Data for Language Documentation, Linguistic Research, and Speech and Language Technologies. Invited keynote presentation at the LDL-2016, 5th Workshop on Linked Data in Linguistics: Managing, Building and Using Linked Language Resources, LREC 2016, Portorož, Slovenia. May 2016.

Endangered Language Documentation: Bootstrapping a Chatino Speech Corpus, Forced Aligner, ASR. LREC 2016, Portorož, Slovenia. Together with Małgorzata E. Ćavar and Hilaria Cruz. May 2016.

Global Open Resources and Information for Language and Linguistic Analysis (GORILLA). LREC 2016, Portorož, Slovenia. Together with Małgorzata E. Ćavar and Lwin Moe. May 2016. See also: GORILLA

Generating a Yiddish Speech Corpus, Forced Aligner and Basic ASR System for the AHEYM Project. LREC 2016, Portorož, Slovenia. Together with Małgorzata E. Ćavar, Dov-Ber Kerler, Anya Quilitsch. May 2016.

The Free Linguistic Environment. Natural Language Processing Seminar, Linguistic Engineering Group at the Institute of Computer Science, Polish Academy of Sciences (ICS PAS), Warsaw, Poland. May, 23rd 2016.

Bootstrapping resources for under-resourced languages: Speech corpora and technologies for the languages of the Balkans. 20th Biennial Conference on Balkan and South Slavic Linguistics. University of Utah, April 2016.

The Free Linguistic Environment (FLE): Theoretical Concepts and Basic Architecture. Together with Lwin Moe. ClingDing, Department of Linguistics, Indiana University, April 2016.

2015

The General Ontology for Linguistic Description (GOLD) and its Role for Digital Language Resources and Language Technology. Workshop "Development of Linguistic Linked Open Data (LLOD) Resources for Collaborative Data-Intensive Research in the Language Sciences," LSA Summer Institute 2015, Chicago, IL. Together with Małgorzata E. Cavar (Indiana). Invited presentation. July 2015.

Documenting Endangered Languages by Developing Speech-Corpora and ASRs: Crow, Hidatsa, Mandan. Together with Małgorzata E. Cavar (Indiana), Wilhelm Meya (TLC), John Boyle (CSU Fresno). Midwest Speech and Language Days (MSLD) 2015, Toyota Technological Institute at Chicago.

Enabling Resources in Digital Language Archives using Automatic Speech Recognition. Together with Małgorzata E. Cavar, Lwin Moe, Aaron Albin. Midwest Speech and Language Days (MSLD) 2015, Toyota Technological Institute at Chicago.

On Split Islands. Department of Linguistics, Indiana University. April 2015.

2014

Automatically Annotated Repository of Digital Video and Audio Resources Community (AARDVARC). Presentation at the ClingDing meeting, Department of Linguistics, Indiana University, together with M.E. Cavar, July 2014.

Morphological grammars and computational analyzer/generators for the documentation of indigenous/endangered languages of the world. Poster presented at the Morphology Fest at Indiana University, Department of Linguistics, together with U. Kazagashewa, M. Mueller, M. Cavar, A. Lamont, S. Fox June 20th, 2014.

Technologies for Bootstrapping of Linguistically Annotated Text Collections with TEI XML Markup. Paper presented at the University of North Texas, Computer Science and Linguistics, together with Malgorzata E. Cavar, February 2014.

The LINGUIST List Now and Then. Paper presented at the University of North Texas, Computer Science and Linguistics, together with Malgorzata E. Cavar, February 2014.

Online visualization of research in historical linguistics. Poster presented at the LSA Annual Meeting 2014 in Minneapolis, MN. Together with M. Cavar, S. Couture, U. Kazagasheva, E. Benzschawel, January, 2014.

2013

Urban fieldwork, evolution, and languages in cities. Paper presented at the Association for the Study of the Arts of the Present ASAP/5: Arts of the City Conference, together with Malgorzata E. Cavar. Wayne State University, October 3rd-6th, 2013.

2012

Bootstrapping large text corpora with TEI XML markup and linguistic annotation. Together with Malgosia E. Cavar. Chicago Colloquium on Digital Humanities and Computer Science. The University of Chicago. 19th of November 2012.

Automatic Linguistic Annotation with TEI-Output. TEI 2012 Conference at Texas A&M, College Station. 9th of November 2012.

The LINGUIST List Corpus: A Large Mailing List Corpus – Management, Annotation and Repository. Together with Malgorzata E. Cavar, Helen Dry Aristar, and Anthony Aristar. TEI 2012 Conference at Texas A&M, College Station. 9th of November 2012.

The Project Gutenberg book archive as a TEI P5 XML text corpus. Together with Malgorzata E. Cavar. TEI 2012 Conference at Texas A&M, College Station. 10th of November 2012.

Dynamic Professional Content Corpora and New Technologies. Wayne State University, 19th of October 2012.

Large Mailing List Corpora: Management, Annotation, and Repository. Together with Helen Aristar-Dry and Anthony Aristar. LREC 2012 Workshop on Challenges in the management of large corpora, Istanbul, 22nd of May, 2012.

Sprachtechnologie und Sprachdokumentation: Eine Darstellung von Projekten des Heimatinstituts der LINGUIST List. Institut für Deutsche Sprache, Mannheim. 8th of May, 2012.

Bootstrapping NLP and MT Resources for under-resourced languages. Cross-lingual Language Technology in service of an integrated multilingual Europe – 20 years on –. University of Hamburg, Germany, 4–5 May

2012.

On Split Islands. Presented at the Syntax/Semantics Discussion Group Meeting in the Linguistics Department at the University of Michigan. 20th of Jan. 2012.

2011

On Split Islands. Presented at the Syntax/Semantics Discussion Group Meeting in the Linguistics Department at the University of Michigan. 20th of Jan. 2012.

Cyclicity and Opacity Effects in the Prosody of Two Different Clitic Classes in New-Shtokavian Variants. Presented at: Ilse Lehiste Memorial Symposium. Together with Malgorzata E. Cavar. 11th of November 2011.

Clitic Placement, Syntactic Discontinuity, and information structure. With Melanie Seiß. LFG 2011, Hong Kong. 17th of July 2011.

The Scheme Natural Language Toolkit (S-NLTK): NLP Library for R6RS and Racket. 4th European Lisp Symposium, Special Focus on Parallelism & Efficiency, TUHH, Hamburg University of Technology, Hamburg, Germany, 1st of April 2011.

No Escape from Clitics in Neo-Shtokavian: Contributions to the syntax and prosody of enclitic auxiliaries and pronouns, Research Colloquium, Linguistics Dept., University of Konstanz, 13th of Jan. 2011.

2010

Some notes on TEI in Linguistics. Plenary presentation at the TEI 2010, the Annual Text Encoding Initiative Conference at the University of Zadar, 13th of November 2010.

Riznica: The Croatian Language Corpus. Joint presentation, presented by Dunja Brozović Rončević, SLAVICORP Workshop, University of Warsaw, Poland, 22nd–24th of November 2010.

O indukciji gramatike: računalni i statistički modeli usvajanja jezika ("On grammar induction: Computational and statistical models of language acquisition"). Research colloquium "*Lingvistička srida*," University of Zadar, 25th of Feb. 2010.

2009

Morfološka analiza i lematizacija hrvatskog jezika na konačnim automatima ("Morphological analysis and lematization of Croatian using Finite State Automata"), Research Colloquium, Zadarska lingvistička Srida, University of Zadar, 19th of November 2009.

Frequency correlations in processing, familiarity, and language usage data of clitics in Croatian. Together with Tomislav Frleta, SLS 2009, 4th Annual Meeting of the Slavic Linguistic Society, 3rd–6th of September 2009.

Empirical evidence for the functional determiner projection in Croatian. Together with Matea Birtić, SLS 2009, 4th Annual Meeting of the Slavic Linguistic Society, 3rd–6th of September 2009.

On bootstrapping of linguistic features for bootstrapping grammars. EACL 2009 workshop Computational Linguistic Aspects of Grammatical Inference, Athens (Greece), 30th of March 2009.

On the induction of linguistic categories and learning grammars. 10. Workshop in Szklarska Poreba (Poland), 12th of March 2009.

Red riječi u hrvatskim govorima: empirijski pristup padežima. Zadarska lingvistička srida, Zadar, 10th of March 2009.

Formalni modeli i Računalna obrada jezika. PhD program in linguistics at the Philosophical Faculty of the University in Osijek. 7th of March 2009.

Nova generacija računalne obrade jezika. Developers User Group Split, 34. Skup IT profesionalaca, Split, Croatia, 22nd of January 2009.

2008

Some Quantitive and Qualitative Aspects of Nominal Case in Croatian. Second Croatian Syntax Days, Osijek, Croatia, 13th–15th of November 2008.

Interoperability and Rapid Bootstrapping of Morphological Parsing and Annotation Automata. IS-LTC 08, Ljubljana, Slovenia. 16th–17th October 2008.

CroMo - Morphological Analysis for Standard Croatian and its Synchronic and Diachronic Dialects and Variants. FSMNLP 2008, Ispra, Lago Maggiore, Italy, 11th–12th of September 2008. (slides in PDF, poster in PDF)

Struktura i razvoj baze podataka za potrebe projekta „Hrvatsko strukovno nazivlje - projekt koordinacije", Poletna Terminološka Šola, ZRC SAZU, Ljubljana, 4th–9th of Sept. 2008.

On Grammar induction, Language Models, and Language Evolution Simulations. PsychoCompLA 2008, at the CogSci 2008, Washington DC, 23rd–24th of July 2008.

Machine learning systems as models for natural language grammar acquisition? University of Nova Gorica, Slovenia. 20th of February 2008.

The Croatian Language Repository: Quantitative and Qualitative Resources for Linguistic Research and Language Technologies. Eastern Michigan University, Institute for Language Information and Technology (ILIT). 31st of January 2008.

The Croatian Language Repository: Quantitative and Qualitative Resources for Linguistic Research and Language Technologies. Indiana University, Linguistics Dept. 29th of January 2008.

2007

Modelle dynamischer Eigenschaften von Sprachen und ihre Anwendungen, University of Bielefeld, Germany, 27th of April 2007.

Dynamische und lernende Sprachmodelle. University of Bochum, Germany, 26th of April 2007.

Dynamic Language Models. JOTA, Ljubljana, Slovenia. 29th of March 2007.

2006

Inducing Lexical Properties with Probabilistic Methods. Polish Academy of Sciences, Warsaw, Poland. 11th of August 2006.

Das Korpus der kroatischen Sprache: Hrvatska jezična mrežna riznica. (The Croatian Language Corpus/Hrvatski jezični korpus). University of Graz, Austria. 19th of June 2006

About Clitics in Croatian: Myths, and Fairy-tales. Together with Dunja Brozović-Rončević, at the conference Hrvatski sintaktički dani (Croatian Syntax Days) at the University of Osijek, Croatia. 11th of May 2006.

Parsing Croatian. Institute of Croatian Language and Linguistics, Zagreb, Croatia. 1st of April 2006.

2005

Inducing Syntactic Properties of Lexical Elements via Information Theoretic Measures. Together with Paul Rodrigues and Giancarlo Schrementi. Workshop on Computational Modeling of Lexical Acquisition: The Split Meeting. Split, Croatia. 25th–28th of July 2005.

Using Morphological and Distributional Cues for Inductive Part-of-Speech Tagging. The Second Midwest Computational Linguistics Colloquium (MCLC-2005), Columbus, OH, The Ohio State University. 15th of May 2005.

Unsupervised morphology induction for part-of-speech tagging. The 29th Penn Linguistics Colloquium. 26th of February 2005.

2004

Constraint-based Cue-learning and cue-based language acquisition. Workshop on Approaches to Empirical Syntax/WOTS-8, Berlin, Germany. 28th of August 2004.

Alignment Based Induction of Morphology Grammar and its Role for Bootstrapping. The 9th conference on Formal Grammar: FGNancy, Nancy, France. 8th of August 2004.

Computationale Modellierung des Spracherwerbs: Wieviel kann man über natürliche Sprache von der Sprache selbst lernen? Center for Computing Technologies (TZI), University of Bremen, Germany. 8th of July 2004.

Syntactic Parsing Using Mutual Information and Relative Entropy. Midwest Computational Linguistics Colloquium (MCLC), Bloomington, Indiana. 25th of June 2004.

Unsupervised Grammar Induction. Cognitive Science Colloquium, Psychology Department, Indiana University, Bloomington, Indiana. 2nd of June 2004.

On unsupervised grammar induction from untagged corpora. Poznań Linguistics Meeting 2004, Poznań, Poland. 19th of May 2004.

Computational Modeling of Language Acquisition. Linguistics Colloquium, University of Potsdam, Germany. 17th of May 2004.

Computational Modeling of Language Acquisition. Psycholinguistics Suppers, CUNY Graduate Center, New York. 4th of May 2004.

Cognitive Aspects of Language Acquisition and Processing. Sveučilište u Splitu (University of Split), Croatian Applied Linguistics Association. 17th of March 2004.

Lexicology and Corpus Linguistics. Sveučilište u Splitu, Anglistika (University of Split, English Department). 16th of March 2004.

2003

Computational Aspects of Language Acquisition. The University of Arizona, Tucson. 4th of March 2003.

Cue-based bootstrapping: Computational Aspects of Language Acquisition. Indiana University, Bloomington. 29th of January 2003.

2002

AME^{WS} – Automatic Metatagging Engine for SemanticWeb with WebService Architecture. University of California (UCLA), Los Angeles. 29th of October 2002.

A Real Live Web Service using Semantic Web Technologies: Automatic Generation of Meta-Information. Conf.: "On The Move Towards Meaningful Internet Systems" (DOA, ODBASE, CoopIS'02). Irvine, California. 28th of October 2002.

AME^{WS} – Automatic Metatagging Engine for SemanticWeb with WebService Architecture. Indiana University, Bloomington: Linguistics Colloquium. 13th of September 2002.

Automatic Generation of Metatags for Intra-Semantic-Web. XSW 2002 Workshop (Semantic Web), Humboldt Universität Berlin. 26th of June 2002.

Zur Rolle von "semantischen Netzen" für das Wissensmanagement ("On the role of semantic nets for knowledge management"). AIK-Symposium on Semantic Web, University of Karlsruhe. 19th of April 2002.

Natürlichsprachliche Systeme für das Wissensmanagement ("Natural language systems for knowledge management"). Colloquium, University of Potsdam. 16th of April 2002.

Distributed Deletion. Kolloqium (Prof. Dr. Günther Grewendorf) University of Frankfurt a. M., Jan. 2002.

2001

Optimizing knowledge mining in the e-back-office. Euromap e2001 conference in Venice. Session Opportunities for language technology in multilingual e-markets. Chairperson: Hanne Fersøe, Center for Sprogteknologi, Denmark. 19th of October 2001.

Digital Dictionary of the 20th Century German Language. Institute of Computer Science, Polish Academy of Sciences. Joint work with Alexander Geyken (BBAW). Warsaw, Poland. 26th of February 2001.

2000

Digital Dictionary of the 20th Century German Language. Jezikoslovne Tehnologije za Slovenski Jezik, JS 2000. Ljubljana, Slovenia. 17th of October 2000.

Klitikpositionen, diskontinuierliche Konstituenten und andere Wortunordnungen im Slavischen. (Clitic positions, discontinuous constituents and other word-diss-orders in Slavic) SlavGG Meeting, University of Leipzig. 15th of July 2000.

1999

Split Constituents. Workshop "Conflicting Rules in Phonology and Syntax", University of Potsdam. Joint work with Gisbert Fanselow. 17th of December 1999.

Discontinuous Constituents in Slavic and Germanic. Drugi Hrvatski Slavistički Kongres (Second Croatian Slavic Congres) in Osijek, Croatia. Joint work with Gisbert Fanselow. 18th of September 1999.

End-to-End Evaluation '99-1. Verbmobil Projektlenkungssitzung, Daimler-Chrysler Center in Stuttgart. 12th of May 1999.

Discontinuous Constituents. Poznań Linguistics Meeting. Poznań, Poland. 1st of May 1999.

1998

End-to-End Evaluation in Verbmobil. Verbmobil Projektlenkungssitzung in Aachen. 7th of December 1998.

Klitike u Hrvatskom Jeziku ("Clitics in Croatian"). Institut za Hrvatski Jezik i Jezikoslovlje (Institute for Croatian language and linguistics), Zagreb, Croatia. 9th of October 1998.

Verbmobil: A Speech-to-Speech Translation System. Institut za Hrvatski Jezik i Jezikoslovlje (Institute for Croatian language and linguistics), Zagreb, Croatia. 8th of October 1998.

Verbmobil: A Speech-to-Speech Translation System. Jezikovne Tehnologije za Slovenski Jezik (JS) '98 (Language technologies for the Slovenian language), Ljubljana, Slovenia. 6th of October 1998.

Autonomy and Look-ahead: Interface phenomena in Croatian. University of Tübingen. 25th of May 1998.

End-to-End Evaluation of Verbmobil. Verbmobil Projektlenkungssitzung, Bonn, Germany. 11th of May 1998.

Split Constituents: A Comparison between Germanic and Slavic. PLM: 31st Poznań Linguistic Meeting (PL). 1st—2nd of May 1998.

1997

Verb Movement and Coordination. FDSL 2, University of Potsdam, joint work with Chris Wilder. 20th—22nd of November 1997.

Interface Phenomena in Slavic: Polish and Croatian Cliticization. FDSL 2, University of Potsdam. 20th–22nd of November 1997.

Children's Sensitivity to Word-Order Variations in German: Evidence for Very Early Parameter Setting. Boston University Conference on Language Development (USA), joint work with: Jürgen Weissenborn, Barbara Höhle, Dorothea Kiefer. 7th–9th of November 1997.

Split Constituents in Germanic and Slavic. International Conference on Pied-Piping, Universität Jena, Germany, joint work with Gisbert Fanselow. 30th of May 1997.

On the Syntactic and Phonological Properties of Enclitic Categories in Slavic. PLM: 30th Poznań Linguistic Meeting (Poland), AG: Reductionist Approaches to Syntax and Descriptive Adequacy (Jacek Witkoś). 3rd of May 1997.

On the Syntax and Phonology of Clitics in Slavic. 19. DGfS Jahrestagung in Düsseldorf, Germany, AG: Die Interaktion grammatischer Teilbereiche (Katharina Hartmann & Daniel Büring). 24th of February 1997.

On Clitics and Split Consituents. Zentrum für allgemeine Sprachwissenschaft (ZAS), Berlin, Germany, Sprachwissenschaftliches Kolloqium. 18th of February 1997.

1996

Interface Phenomena in Croatian and Polish. University of Leipzig, Germany. 27th of November 1996.

Frühe Syntaktische Wissensstrukturen: Kontinuität und Ökonomie ("Early syntactic knowledge structures: continuity and economy"). Annual meeting of the RULE group, Groß Dölln, Germany, (PANNONIA Hotel Döllnsee), Berlin-Brandenburg Academy of Science. 11th of October 1996.

Frühe Syntaktische Wissensstrukturen: Kontinuität und Ökonomie (Early syntactic knowledge structures: continuity and economy). Workshop: TROPICS, Berlin-Brandenburgische Akademie der Wissenschaften,

joint work with: Jürgen Weissenborn, Barbara Höhle, Dorothea Kiefer. 28th of September 1996.

On the Structure of Tacit Syntactic Knowledge: Continuity and Economy. With Weissenborn, J., Höhle, B., Kiefer, D., Conference: Approaches to Bootstrapping in Early Language Development, Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, 1996.

Facing the Interface: On the Properties of "Deficient Elements" in South-Slavic. Third Summer School in Generative Linguistics, Palacky Univerzitet Olomouc (Czech Republic). 22nd of August 1996.

On Clitics in Croatian: More Syntax than Prosody! Geisteswissenschaftliche Zentren Berlin e.V., Zentrum für allgemeine Sprachwissenschaft, Typologie und Universalienforschung (ZAS), Workshop on the Syntax, Morphology and Phonology of Clitics. 26th–27th of May 1996.

1995

Auxiliaries in Serbian/Croatian and English. FDSL 1, University of Leipzig (D). Dec. 1995.

1993

Minimalistische Bewegung (Minimalistic Movement). University of Regensburg (D), 15th of December 1993.

Triggers & Economy. Universität Wuppertal (D). 12th of November 1993.

Pronominal and Verbal Clitics in Croatian. University of Durham (UK), EUROTYP Clitics Meeting. 17th of October 1993.

Pronominal and Verbal Clitics in Croatian. SOAS, London (UK) (School of Oriental and African Studies). 14th of October 1993.

Verbale und pronominale Klitika (Verbal and pronominal clitics). Universität zu Köln (D), GGS. 10th of July 1993.

Ein Konzept für MultiMedia im Sprachunterricht am Beispiel des Japanischen ("A concept for multimedia in 2nd language teaching on the basis of Japanese"). München (D), MediaNet, Universität Frankfurt a.M. 9th of July 1993.

Wortstellungsvariation, Verbbewegung und Ökonomie-Prinzipien ("Word order variation, verb movement, and economy principles"). Universität Stuttgart (D). 24th of June 1993.

Verbs and Clitics in Croatian. Université de Genève (CH), Clitic Workshop. 1st of June 1993.

Verbs and Clitics in Croatian. Rijksuniversitet Groningen (NL). 28th of May 1993.

X^0-Bewegung und Ökonomie (X^0-movement and economy). DGfS annual meeting, University of Jena (D), Workshop: Wortordnung (word order). 3rd of March 1993.

1992

X^0-Bewegung und Ökonomie (X^0–movement and economy). Max-Planck-Gesellschaft zur Förderung der Wissenschaften e.V., Arbeitsgruppe Strukturelle Grammatik an der Humboldt-Universität, Berlin. 17th of November 1992.

Long Head Movement? Verb-Movement and Cliticization in Croatian. University of Regensburg, Germany, GGS. 1st of February 1992.

Teaching

Indiana University

Academic year 2024–2025:

Academic year 2023–2024:

Academic year 2022–2023:

LING 645 / CSCI-B 659 Topics in Artificial Intelligence – Advanced Natural Language Processing. Fall 2023. 3 credits graduate level course, 28 x 1.5 hours.

CSCI-B 659 Topics in Artificial Intelligence – Large Language Models and Knowledge Graphs (LING 715). Fall 2023. 3 credits graduate level course, 28 x 1.5 hours.

Academic year 2022–2023:

sabbatical leave, NLP-Lab research group administration

Academic year 2021–2022:

no teaching, NLP-Lab research group administration

Academic year 2020–2021:

LING 645 / CSCI-B 659 Topics in Artificial Intelligence – Advanced Natural Language Processing. Fall 2020. 3 credits graduate level course, 28 x 1.5 hours.

LING 615 Corpus Linguistics. Fall 2020. 3 credits graduate level course, 28 x 1.5 hours.

Academic year 2020–2021:

LING 645 / CSCI-B 659 Topics in Artificial Intelligence – Advanced Natural Language Processing. Fall 2020. 3 credits graduate level course, 28 x 1.5 hours.

LING 615 Corpus Linguistics. Fall 2020. 3 credits graduate level course, 28 x 1.5 hours.

Academic year 2019–2020:

CSCI-B 659 Topics in Artificial Intelligence – Deep Learning for NLP (LING 665). Spring 2020. 3 credits graduate level course, 28 x 1.5 hours.

LING 614 Alternative Syntactic Theories. Spring 2020. 3 credits graduate level course, 28 x 1.5 hours.

CSCI-B 659 Topics in Artificial Intelligence – Advanced Natural Language Processing. Fall 2019. 3 credits graduate level course, 28 x 1.5 hours.

LING 615 Corpus Linguistics. Fall 2019. 3 credits graduate level course, 28 x 1.5 hours.

Academic year 2018–2019:

CSCI-B 659 Topics in Artificial Intelligence – Deep Learning for NLP (LING 665). Spring 2019. 3 credits graduate level course, ca. 44 students, 28 x 1.5 hours.

LING 614 Alternative Syntactic Theories. Spring 2019. 3 credits graduate level course, 8 students, 28 x 1.5 hours.

CSCI-B 659 Topics in Artificial Intelligence – Advanced Natural Language Processing. Fall 2018. 3 credits graduate level course, ca. 40 students, 28 x 1.5 hours.

Academic year 2017–2018:

CSCI-B 659 Topics in Artificial Intelligence – Deep Learning for NLP (LING 665). Spring 2018. 3 credits graduate level course, ca. 40 students, 28 x 1.5 hours.

CSCI-B 659 Topics in Artificial Intelligence – Advanced Natural Language Processing. Fall 2016. 3 credits graduate level course, ca. 40 students, 28 x 1.5 hours.

CSCI-B 659 Topics in Artificial Intelligence: Semantic Natural Language Processing, NoAI, and Big Knowledge (including dialog and AI) (LING 715). Fall 2017. 3 credits graduate level course, 30 students, 28 x 1.5 hours.

Academic year 2016–2017:

LING 645 Advanced Natural Language Processing / CSCI-B 659 Topics in Artificial Intelligence. Fall 2016. 3 credits graduate level course, 47 students, 28 x 1.5 hours.

LING 614 Alternative Syntactic Theories. Spring 2017. 3 credits graduate level course, 10 students, 28 x 1.5 hours.

LING 665 and CSCI-B 659: Applying Machine Learning Techniques in Computational Linguistics. Spring 2017. 3 credits graduate level course, ca. 40 students, 28 x 1.5 hours.

International Summer Schools and Guest-lectures

Summer 1995:

Introduction to Language Processing. 2nd Central European Summer School in Generative Syntax, Palacký University in Olomouc (Czech Republic). 5 x 2 hours, `http://egg.auf.net/95/`

June 2004:

Computational Modeling. Summer school "Syntaxfest", MA and PhD level course, Indiana University. 5 x 2 hours

December 2004:

Introduction to Computational Modeling of Lexical and Grammatical Knowledge Acquisition using Machine Learning Techniques. Block seminar, University of Potsdam, Institute of Linguistics.

September 2005:

Python für die Computerlinguistik. MA and PhD students, DGfS Fall School at the University in Bochum. 5 x 2 hours, `http://ling.unizd.hr/~dcavar/pycl/`

Summer 2006:

Introduction to symbolic and statistical NLP in Scheme. MA and PhD level course, ESSLLI 18th European Summer School in Logic Language and Information. University of Malaga, Spain. 5 x 2 hours, `http://esslli2006.lcc.uma.es/give-page.php?id=6`, `http://ling.unizd.hr/~dcavar/ESSLLI2006/`

Introduction to Scheme/Python for Computational Linguistics. MA and PhD level course, Jadertina Summer School in Empirical and Computational Linguistics, University in Zadar, Croatia. 5 x 2 hours, `http://ling.unizd.hr/~dcavar/JSSECL2006Course/`

Summer 2013:

Python for Linguists, LSA Summer Institute at the University of Michigan, Ann Arbor (`http://lsa2013.lsa.umich.edu/`). 5 x 2 hours.

Summer 2024:

Damir Cavar and Billy Dickson: *Generative AI and Symbolic Knowledge Representations: Large Language Models, Knowledge, and Reasoning*, European Summer School in Logic, Language and Information (ESSLLI), Leuven, Belgium (`https://2024.esslli.eu/placeholder-programme/course-schedule.html`). 5 x 1.5 hours + student advising.

Extracurricular Courses and Classes

Indiana University 2014–2018:

- *Introduction to Haskell for NLP* (multiple sessions during the summer months, summer 2016).
- *Meta-data standards for NLP Data* (3 x 2 hours, 2016).

- *C++ implementation of a Probabilistic LFG morphological analyzer and parser* (weekly about 2 hours for almost 2 years, about 10 students participated).
- *Linear Algebra for Machine Learning, Deep Learning, NLP* (summer and fall 2017, weekly about 2 hours, about 4 students participated during the summer).
- *Speech Processing for Supra-segmental analysis in Computational Pragmatics and Semantics, using Deep Learning approaches* (summer 2018, weekly for 2 hours with approx. 6 students)
- *Open Information Extraction, Knowledge Graphs, OWL and reasoning using Graph-DBs and Deep Learning* (since summer 2016, weekly for 2 hours, joint group of students from Kelly and COAS, a group of more than 10 students was involved, summer 2018 it is now about 15 students from SICE and COAS)
- *Quantum Algorithms using Linear Algebra* (summer 2018, weekly for 2 hours, about 10 students and multiple colleagues participating)

University of Zadar 2010:

- *Schemers in Zadar* (Scheme programming and S-NLTK) weekly meetings of 2 hours, approx. 40, University of Zadar.

Last updated: April 12, 2025