Marcel **Bollmann**

Associate Professor in Computer Science and Natural Language Processing

+46 (0)13-28 1572 | marcel.bollmann@liu.se | marcel.bollmann.me | mbollmann

Personal Details

Born 5 September 1984
Citizenship German

Academic Website https://marcel.bollmann.me/
Google Scholar https://scholar.google.com/citations?user=l3pm9QkAAAAJ
Semantic Scholar https://www.semanticscholar.org/author/Marcel-Bollmann/34887843
ORCID iD https://orcid.org/0000-0003-2598-8150

Research Interests

- Natural language processing (NLP), particularly for low-resource languages, morphologically rich languages, and non-standard varieties
- Machine learning, neural networks & deep learning, representation learning
- Multilinguality & linguistically-informed approaches for NLP

Research Experience

Linköping University, Department of Computer and Information Science (IDA)

Linköping, Sweden

Associate Professor (Universitetslektor) in Computer Science

since 16.01.2023

- Member of the Natural Language Processing Group at IDA/AIICS
- Research and teaching in computer science and natural language processing

Jönköping University, School of Engineering, Department of Computing

Jönköping, Sweden

Assistant Professor (Universitetslektor) in Computer Science

01.06.2021 – 15.01.2023

- Member of the Jönköping Artificial Intelligence Lab (JAIL)
- Teaching in data science, AI, and natural language processing

University of Copenhagen, Department of Computer Science (Prof. Anders Søgaard)

Copenhagen, Denmark

Postdoctoral Researcher & Marie Skłodowska-Curie Fellow

01.04.2019 – 31.03.2021

- Project on "Morphologically-informed representations for natural language processing" (MorphIRe)
- Received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 845995.

Postdoctoral Researcher

01.01.2018 – 31.12.2018

- Deep learning for NLP, incl. cross-lingual and multi-task learning for low-resource and historical language.

Ruhr-Universität Bochum, Department of Linguistics (Prof. Stefanie Dipper)

Bochum, Germany

Research Assistant, part-time (65%)

01.03.2011 – 31.08.2017

- Researcher in projects related to computational historical linguistics (https://www.linguistics.rub.de/comphist/)
  - "St. Anselmi Fragen an Maria" ('Questions by Saint Anselm to the Virgin Mary')
  - "Reference Corpus Early New High German"
  - "Reference Corpus Middle High German"
- Worked on supervised machine learning for historical text normalization, developing new algorithms and experimenting with state-of-the-art neural network architectures.
- Developed open-source software tool for text normalization (Norma: https://github.com/comphist/norma).
- Developed open-source software tool for web-based linguistic annotation (CorA: https://github.com/comphist/cora) in close collaboration with linguists and German philologists.

Education

Ruhr-Universität Bochum

Bochum, Germany

Dr. phil. in Computational Linguistics

20.06.2018

- Final grade: **'summa cum laude'** *(highest possible grade)*
- Dissertation topic: Normalization of Historical Texts with Neural Network Models

M.A. in Linguistics (with focus on Computational Linguistics) *07.12.2012*

- Final grade: **'with distinction'** *(highest possible grade)*
- Thesis topic: Automatic normalization for linguistic annotation of historical language data
- **Won an award** for best student thesis 2011–2013 by the GSCL (German Society for Computational Linguistics & Language Technology).

B.A. in Mathematics and Linguistics (with focus on Computational Linguistics) *09.07.2009*

- Thesis topic: Comparing semantic distance measures on WordNet and GermaNet

Awards & Grants

2021 **Best Paper Award**, Conference of the European Chapter of the Association for Computational Linguistics (EACL)

2021 **Outstanding Reviewer**, Conference of the European Chapter of the Association for Computational Linguistics (EACL)

2019 **Research Grant**, Marie Skłodowska-Curie Individual Fellowship (MSCA-IF), Grant No. 845995

- Title of project: "Morphologically-informed representations for natural language processing — MorphIRe" (https://cordis.europa.eu/project/rcn/221604/factsheet/en)
- Grant amount: € 207,312
- Proposal evaluated with score of 100% (top 0.55% in "Information Science and Engineering" panel, top 0.09% overall)

2018 **Nvidia GPU Grant**, Nvidia Corporation, donation of a GeForce Titan Xp

2017 **Travel Grant**, German Academic Exchange Service (DAAD), for presenting at the ACL conference in Vancouver, Canada

2013 **Best Student Thesis Award**, German Society for Computational Linguistics & Language Technology (GSCL)

Teaching

Jönköping University, Department of Computing *Jönköping, Sweden*

Course Developer, Coordinator, and Teacher *2021 – 2022*

- "Course developer" includes full preparation of teaching materials.

2022 **Course Developer & Coordinator**, "Natural Language Processing"

2022 **Course Developer & Coordinator**, "Data Science Programming" using Python and R

2022 **Course Coordinator**, "State of the Art in AI Research"

2021 **Course Coordinator**, "State of the Art in AI Research"

Lecturer, Tutor, Seminar Chair

2021 **Tutor**, "Programmeringsteknik" (Introduction to Programming in C)

2021 **Lecturer**, "Data visualization" and "Experimental design" as part of "Research Methods" course

University of Copenhagen, Department of Computer Science *Copenhagen, Denmark*

Lecturer *2018 – 2019*

2019 **Lecturer**, "Sequence labelling" and "Machine translation" as part of "Natural Language Processing" course

2018 **Lecturer**, "Sequence labelling" as part of "Natural Language Processing" course

Ruhr-Universität Bochum, Department of Linguistics *Bochum, Germany*

Lecturer, Tutor, Seminar Chair *2008 – 2015*

- All seminars & tutorials involved full preparation of teaching materials.
- Seminars included individual supervision and grading of students' coursework.

2015 **Lecturer & Seminar Chair**, "Aspects of natural language generation"

2014 **Lecturer & Seminar Chair**, "Non-standard language data"

2013 **Lecturer & Seminar Chair**, "Aspects of natural language generation"

2010 **Tutorial**, "Tools & Techniques" for computational linguistics (3x)

Formal qualifications

2021 **Introduction to University Pedagogy (3 CP)**, University of Copenhagen, Denmark

Talks & Presentations

Conference presentations

Talks EACL 2021, ACL 2020, ACL 2017, LAW 2016, COLING 2016, LAW 2013, LTC 2011, + 3 further workshops

Posters ACL 2019, NAACL 2019, DeepLo 2018, CLIN 2017, LaTeCH 2014, DGfS 2012, ENLG 2011

Invited talks

2022 **Seminar at Department of Computer and Information Science, Linköping University**, Linköping, Sweden
- Invited talk on "From historical texts to creoles: NLP for challenging domains" (all expenses paid)

2022 **AI@JKPG (Company Network Event)**, Jönköping, Sweden
- Invited talk on "State of the art in natural language processing" (company networking event)

2018 **Almanach Seminar Series at Inria**, Paris, France
- Invited talk on "Historical text normalization with neural networks" (all expenses paid)

2015 **Lecture Series at the Department of Linguistics**, Bochum, Germany
- Invited talk on "Automatic annotation of historical language data"

Outreach

2020 **research*eu magazine**, European Commission
- Article "Equity in natural language processing" about my "MorphIRe" project (https://cordis.europa.eu/article/id/421665-equity-in-natural-language-processing)

Volunteer & Community Work

2019– **Site Development Lead**, ACL Anthology (https://aclweb.org/anthology/)
- Main archive of research papers in computational linguistics & natural language processing
- Full re-implementation and partial re-design of the website between 01/2019 and 03/2019; work performed "to extremely high standards" (https://www.aclweb.org/adminwiki/index.php?title=2019Q1_Reports:_Anthology_Director)

2016 **Local Organizer**, KONVENS 2016 (Conference on Natural Language Processing), Bochum, Germany
- Implementation & design of conference handbook & website (https://www.linguistics.rub.de/konvens16/)

Reviewing & Program Committees

Journals Computational Linguistics, Language Resources and Evaluation (LREV), Natural Language Engineering (NLE), Information Processing and Management

Conferences ACL 2019– (incl. ACL Rolling Review), COLING 2018–2022, EACL 2021, EMNLP 2020, ICCC 2022, IJCAI 2021 & 2022, KONVENS 2018–2022, NAACL-HLT 2018–2019, NoDaLiDa 2019–2021

Workshops Language Technologies for Historical and Ancient Languages (LT4HALA) 2020–2022, Universal Dependencies Workshop (UDW) 2020, NLP Open Source Software (NLP-OSS) 2020

Grants Austrian Science Fund (FWF)

Skills

Natural Languages

German	*(native)*	●●●●●
English	*(fluent)*	●●●●●
Swedish	*(intermediate)*	●●●○○
Danish	*(basic)*	●●○○○
Japanese	*(elementary)*	●○○○○

Programming Languages

Python	●●●●●
Bash, R	●●●●○
JavaScript, PHP, HTML, CSS	●●●●○
C, C++, Java, Rust	●●○○○
Perl, Ruby	●○○○○

IT/Software

General Windows, Linux (incl. administration & command-line tools), LaTeX

Scientific Development Machine learning frameworks (*PyTorch, Keras, Tensorflow, DyNet*); Data visualization (*Seaborn, matplotlib*)

Version control (*Git*), databases (*MySQL, SQLite*), testing & documentation frameworks, continuous integration (*Travis*), static website generators (*Hugo, webgen*)

Teaching Online learning management platforms (*Canvas, Moodle, Blackboard*), interactive computing (*Jupyter Notebook*), online assessment (*Inspera*), student administration (*Ladok*)

Publications

- **Google h-index: 13** (https://scholar.google.com/citations?user=l3pm9QkAAAAJ)

Conferences (all peer-reviewed)

- Marcel Bollmann and Anders Søgaard (2021). "Error Analysis and the Role of Morphology". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1887–1900. URL: https://www.aclweb.org/anthology/2021.eacl-main.162
  - Trained a classifier to predict errors made by NLP models, using only morphological features as input.
  - Won a **Best Paper Award** at the conference.
- Marcel Bollmann and Desmond Elliott (2020). "On Forgetting to Cite Older Papers: An Analysis of the ACL Anthology". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7819–7827. DOI: 10.18653/v1/2020.acl-main.699. URL: https://www.aclweb.org/anthology/2020.acl-main.699
  - Performed a statistical analysis of outgoing citations in NLP research papers.
- Marcel Bollmann (2019). "A Large-Scale Comparison of Historical Text Normalization Systems". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 3885–3898. URL: http://www.aclweb.org/anthology/N19-1389
  - Surveyed, evaluated, and thoroughly analyzed methods for automatic normalization of historical texts on corpora from eight languages. Based on my PhD work.
- Meriem Beloucif, Ana Valeria Gonzalez, Marcel Bollmann, and Anders Søgaard (2019). "Naive Regularizers for Low-Resource Neural Machine Translation". In: *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Varna, Bulgaria: INCOMA Ltd., pp. 102–111. DOI: 10.26615/978-954-452-056-4_013. URL: https://www.aclweb.org/anthology/R19-1013
  - Contributed to implementation and evaluation of a low-resource machine translation model.
- Simon Flachs, Marcel Bollmann, and Anders Søgaard (2019). "Historical Text Normalization with Delayed Rewards". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 1614–1619. URL: https://www.aclweb.org/anthology/P19-1157
  - Applied reinforcement learning techniques to historical text normalization.
- Marcel Bollmann, Joachim Bingel, and Anders Søgaard (2017). "Learning Attention for Historical Text Normalization by Learning to Pronounce". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 332–344. DOI: 10.18653/v1/P17-1031. URL: https://www.aclweb.org/anthology/P17-1031
  - Applied neural encoder–decoder models to historical text normalization, using multi-task learning to simultaneously train them on learning pronunciation.
- Marcel Bollmann and Anders Søgaard (2016). "Improving Historical Spelling Normalization with Bi-Directional LSTMs and Multi-Task Learning". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*. Osaka, Japan. URL: https://aclweb.org/anthology/C16-1013
  - Applied a deep neural network model using bidirectional long short-term memory (bi-LSTM) units to historical text normalization, showing that multi-task learning can improve the model when training in low-resource scenarios.

Journals (all peer-reviewed)

- Erik Tjong Kim Sang, Marcel Bollmann, Remko Boschker, Francisco Casacuberta, Feike Dietz, Stefanie Dipper, Miguel Domingo, Rob van der Goot, Marjo van Koppen, Nikola Ljubešić, Robert Östling, Florian Petran, Eva Pettersson, Yves Scherrer, Marijn Schraagen, Leen Sevens, Jörg Tiedemann, Tom Vanallemeersch, and Kalliopi Zervanou (2017). "The CLIN27 Shared Task: Translating Historical Text to Contemporary Language for Improving Automatic Linguistic Annotation". In: *Computational Linguistics in the Netherlands Journal* 7, pp. 53–64. URL: https://clinjournal.org/clinj/article/view/68/61
  - Participated in a shared task on historical text normalization, achieving a shared 1st rank using a character-based neural network model.
- Florian Petran, Marcel Bollmann, Stefanie Dipper, and Thomas Klein (2016). "ReM: A Reference Corpus of Middle High German — Corpus Compilation, Annotation, and Access". In: *Journal for Language Technology and Computational Linguistics (JLCL)* 31.2. Ed. by Armin Hoenen, Alexander Mehler, and Jost Gippert, pp. 1–15. URL: https://jlcl.org/content/2-allissues/3-Heft2-2016/01Petran.pdf
  - Developed semi-automatic annotation tools and an XML-based data format for a corpus of historical German documents, made publicly available at https://www.linguistics.rub.de/rem/.

- Marcel Bollmann, Florian Petran, and Stefanie Dipper (2014). "Applying Rule-Based Normalization to Different Types of Historical Texts—An Evaluation". In: *Human Language Technology. Challenges for Computer Science and Linguistics. 5th Language and Technology Conference, LTC 2011. Revised Selected Papers*. Ed. by Zygmunt Vetulani and Joseph Mariani. 1st. Lecture Notes in Artificial Intelligence 8387. Springer International Publishing, pp. 166–177. DOI: 10.1007/978-3-319-08958-4

  – Developed a novel rule-based algorithm for historical text normalization in low-resource scenarios.

Workshops (all peer-reviewed)

- Rahul Aralikatte, Héctor Ricardo Murrieta Bello, Miryam de Lhoneux, Daniel Hershcovich, Marcel Bollmann, and Anders Søgaard (2021). "How far can we get with one GPU in 100 hours? CoAStaL at MultiIndicMT Shared Task". In: *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*. Online: Association for Computational Linguistics, pp. 205–211. DOI: 10.18653/v1/2021.wat-1.24. URL: https://aclanthology.org/2021.wat-1.24
- Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard (2021). "Moses and the Character-Based Random Babbling Baseline: CoAStaL at AmericasNLP 2021 Shared Task". In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Online: Association for Computational Linguistics, pp. 248–254. DOI: 10.18653/v1/2021.americasnlp-1.28. URL: https://aclanthology.org/2021.americasnlp-1.28
- Marcel Bollmann, Natalia Korchagina, and Anders Søgaard (2019). "Few-Shot and Zero-Shot Learning for Historical Text Normalization". In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, pp. 104–114. DOI: 10.18653/v1/D19-6112. URL: https://www.aclweb.org/anthology/D19-6112
- Marcel Bollmann, Anders Søgaard, and Joachim Bingel (2018). "Multi-Task Learning for Historical Text Normalization: Size Matters". In: *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*. Association for Computational Linguistics, pp. 19–24. URL: https://aclweb.org/anthology/W18-3403
- Marcel Bollmann, Stefanie Dipper, and Florian Petran (2016). "Evaluating Inter-Annotator Agreement on Historical Spelling Normalization". In: *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*. Berlin, Germany: Association for Computational Linguistics, pp. 89–98. URL: https://aclweb.org/anthology/W16-1711
- Marcel Bollmann, Florian Petran, Stefanie Dipper, and Julia Krasselt (2014). "CorA: A Web-Based Annotation Tool for Historical and Other Non-Standard Language Data". In: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*. Gothenburg, Sweden, pp. 86–90. URL: https://aclweb.org/anthology/W14-0612
- Marcel Bollmann (2013b). "POS Tagging for Historical Texts with Sparse Training Data". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability in Discourse*. Sofia, Bulgaria, pp. 11–18. URL: https://aclweb.org/anthology/W13-2302
- Marcel Bollmann (2012). "(Semi-)Automatic Normalization of Historical Texts Using Distance Measures and the Norma Tool". In: *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*. Lisbon, Portugal. URL: https://marcel.bollmann.me/pub/acrh12.pdf
- Marcel Bollmann, Stefanie Dipper, Julia Krasselt, and Florian Petran (2012). "Manual and Semi-Automatic Normalization of Historical Spelling – Case Studies from Early New High German". In: *LThist 2012: First International Workshop on Language Technology for Historical Text(s)*. Vienna, Austria, pp. 342–350. URL: https://marcel.bollmann.me/pub/lthist12.pdf
- Marcel Bollmann, Florian Petran, and Stefanie Dipper (2011). "Rule-Based Normalization of Historical Texts". In: *Proceedings of the International Workshop on Language Technologies for Digital Humanities and Cultural Heritage*. Hissar, Bulgaria, pp. 34–42. URL: https://marcel.bollmann.me/pub/ranlp11.pdf
- Marcel Bollmann (2011). "Adapting SimpleNLG to German". In: *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG 2011)*. Nancy, France, pp. 133–138. URL: https://aclweb.org/anthology/W11-2817

Theses & Reports

- Marcel Bollmann (2018). "Normalization of Historical Texts with Neural Network Models". In: *Bochumer Linguistische Arbeitsberichte* 22. Revised and updated version of PhD thesis. URL: https://www.linguistics.rub.de/forschung/arbeitsberichte/22.pdf

  – Revised and updated version of my PhD thesis.

- Julia Krasselt, Marcel Bollmann, Stefanie Dipper, and Florian Petran (2015). "Guidelines für die Normalisierung historischer deutscher Texte / Guidelines for Normalizing Historical German Texts". In: *Bochumer Linguistische Arbeitsberichte* 15. URL: https://www.linguistics.rub.de/forschung/arbeitsberichte/15.pdf
- Marcel Bollmann (2013a). "Automatic Normalization for Linguistic Annotation of Historical Language Data". In: *Bochumer Linguistische Arbeitsberichte* 13. Revised version of M.A. thesis. URL: https://www.linguistics.rub.de/forschung/arbeitsberichte/13.pdf

  – Revised and updated version of my M.A. thesis.