

Marcel Bollmann

POSTDOCTORAL RESEARCHER, COMPUTATIONAL LINGUIST

Frankrigsgade 19, 4.tv · 2300 København S · Denmark

☎ (+45) 5357-5984 | ✉ marcel@di.ku.dk | 🏠 marcel.bollmann.me | 🌐 mbollmann

Personal Details

Born 5 September 1984
Citizenship German

Personal Website <https://marcel.bollmann.me/>
Google Scholar Profile <https://scholar.google.com/citations?user=l3pm9QkAAAAJ>

Research Interests

- Natural language processing (NLP), particularly for low-resource languages, morphologically rich languages, and non-standard varieties
- Typologically and morphologically-informed representations for NLP models
- Machine learning, neural networks, deep learning, multi-task learning

Research Experience

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)
- Receives 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

- Focus on deep learning for NLP, particularly 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 creating and processing historical corpora (<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 automatic spelling normalization algorithms for historical language data.
- 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>).

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

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

University of Copenhagen, Department of Computer Science

Copenhagen, Denmark

LECTURER

2018

- 2018 **Guest Lecturer**, Session on “Sequence labelling” as part of “Natural Language Processing” course

Ruhr-Universität Bochum, Department of Linguistics

Bochum, Germany

LECTURER/TUTOR

2008 – 2015

- All courses involved full preparation of teaching materials from scratch.
- Seminars included supervision and grading of students’ coursework.

- 2015 **Seminar**, Aspects of natural language generation
- 2014 **Seminar**, Non-standard language data
- 2013 **Seminar**, Aspects of natural language generation
- 2010 **Tutorial**, “Tools & Techniques” for computational linguistics
- 2009 **Tutorial**, “Tools & Techniques” for computational linguistics
- 2008 **Tutorial**, “Tools & Techniques” for computational linguistics

Talks

INVITED TALKS

- 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”

ORAL CONFERENCE PRESENTATIONS

- 2017 **Annual Meeting of the Association for Computational Linguistics (ACL)**, Vancouver, Canada
- 2016 **Linguistic Annotation Workshop (LAW)**, Berlin, Germany
- 2016 **International Conference on Computational Linguistics (COLING)**, Osaka, Japan
- 2013 **Workshop on Linguistic Annotation and Interoperability in Discourse (LAW & ID)**, Sofia, Bulgaria
- 2013 **Workshop on Corpus Analysis with Noise in the Signal (CANS)**, Lancaster, Great Britain
- 2012 **Workshop on Language Technology for Historical Text(s) (LTHist)**, Vienna, Austria
- 2012 **Workshop on Annotation of Corpora for Research in the Humanities (ACRH)**, Lisbon, Portugal
- 2011 **Language & Technology Conference (LTC)**, Poznań, Poland

Volunteer Work

- 2019 **Site Development Lead**, ACL Anthology (<https://aclweb.org/anthology/>)
- 2016 **Local Organizer**, KONVENS 2016 (Conference on Natural Language Processing), Bochum, Germany

REVIEWING & PROGRAM COMMITTEES

- Journals** Computational Linguistics, Language Resources and Evaluation (LREV), Natural Language Engineering (NLE), Information Processing and Management
- Conferences** ACL 2019, NAACL-HLT 2019 & 2018, NoDaLiDa 2019, COLING 2018, KONVENS 2019 & 2018
- Workshops** NoDaLiDa 2017 Workshop on Processing Historical Language

Publications

- Google h-index: 9

CONFERENCES (ALL PEER-REVIEWED)

- Marcel Bollmann (2019). “A Large-Scale Comparison of Historical Text Normalization Systems”. In: *arXiv:1904.02036 [cs]*. To appear at NAACL 2019. URL: <https://arxiv.org/pdf/1904.02036>
 - Surveyed and evaluated methods for automatic normalization of historical texts on corpora from eight languages. Based on my PhD work.
- 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](https://doi.org/10.18653/v1/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: <http://www.clinjournal.org/sites/clinjournal.org/files/04.clin27-shared-task.pdf>
 - 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://j1cl.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 (2015). “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 Mariana. 1st. Lecture Notes in Artificial Intelligence 8387. Springer, pp. 166–177. doi: [10.1007/978-3-319-08958-4](https://doi.org/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)

- 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, Natalia Korchagina, and Anders Søgaard (2019). “Few-Shot and Zero-Shot Learning for Historical Text Normalization”. In: *arXiv:1903.04870 [cs]*. URL: <https://arxiv.org/pdf/1903.04870>
- 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.

Skills

NATURAL LANGUAGES

German	<i>(native)</i>	●●●●●●
English	<i>(fluent)</i>	●●●●●●
Danish	<i>(basic)</i>	●●●●●●
French	<i>(elementary)</i>	●●●●●●
Japanese	<i>(elementary)</i>	●●●●●●

PROGRAMMING LANGUAGES

Python	●●●●●●
Shell Scripting (bash, fish)	●●●●●●
JavaScript, PHP, HTML, CSS	●●●●●●
C++, Java, Rust	●●●●●●
Perl, Ruby	●●●●●●

IT/SOFTWARE

General	Windows, Linux (incl. administration & command-line tools), \LaTeX
Scientific	Machine learning frameworks: Keras, Tensorflow, DyNet, CRFsuite
Development	Databases (MySQL/SQLite), version control (SVN/Git), testing & documentation frameworks, continuous integration (Travis)