Marcel **Bollmann**

Postdoctoral Researcher, Computational Linguist

Tietgensgade 68, 5.tv · 1704 København V · 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 — *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 (Expected) — *20.06.2018*

- Final grade: **'summa cum laude'** *(highest possible grade)*
- Dissertation topic: Normalization of Historical Texts with Neural Network Models
- *Note: All exams passed; formal completion only after fulfilling publication requirement, expected November 2018*

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

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*

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

Committees & Reviewing

2016 **Local Organizer**, KONVENS 2016 (Conference on Natural Language Processing), Bochum, Germany

Reviewing

Journals Computational Linguistics, Language Resources and Evaluation (LREV), Natural Language Engineering (NLE)
Conferences NAACL HLT 2018, COLING 2018, KONVENS 2018
Workshops NoDaLiDa 2017 Workshop on Processing Historical Language

Skills

Natural Languages

German *(native)*
English *(fluent)*
Danish *(basic)*
French *(elementary)*
Japanese *(elementary)*

Programming Languages

Python
Shell Scripting (bash, fish)
JavaScript, PHP, HTML, CSS
C++, Java
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)

Publications

- **Google h-index: 9**

Conferences (all peer-reviewed)

- 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)*. Vancouver, Canada: Association for Computational Linguistics, pp. 332–344. DOI: 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: http://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: http://www.jlcl.org/2016_Heft2/Heft2-2016.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. URL: http://link.springer.com/book/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*. Melbourne: Association for Computational Linguistics, pp. 19–24. URL: http://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: http://anthology.aclweb.org/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: http://aclweb.org/anthology//W/W14/W14-0612.pdf

- 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: http://aclweb.org/anthology/W/W13/W13-2302.pdf

- 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: *Proceedings of the First International Workshop on Language Technology for Historical Text(s) (LThist 2012)*. 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: http://www.aclweb.org/anthology/W11-2817

Theses & Reports

- Marcel Bollmann (to appear). “Normalization of Historical Texts with Neural Network Models”. In: *Bochumer Linguistische Arbeitsberichte*

  — 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: http://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. URL: http://www.linguistics.rub.de/forschung/arbeitsberichte/13.pdf

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