Curriculum Vitae

Dr. phil. Bryan Jurish

Amselhainstraße 86 · 14612 Falkensee · Germany moocow@cudmuncher.de

PERSONAL INFORMATION

Born 4th March, 1974 in Hinsdale, IL, USA; married December 2005; 2 children.

Resident in Germany since July, 1996 (*Niederlassungserlaubnis* issued September, 2005).

PROFESSIONAL EXPERIENCE

Senior Computational Linguist, 2txt – natural language generation GmbH 2021 – present

Research Associate, Berlin-Brandenburgische Akademie der Wissenschaften 2005 – 2021

Research Associate, Universität Potsdam 2003 – 2005

Research Fellow, Berlin-Brandenburgische Akademie der Wissenschaften 2002 – 2003

Student Assistant, Berlin-Brandenburgische Akademie der Wissenschaften 2001 – 2002

Student Assistant, Computational Linguistics, Universität Potsdam 1997 – 2000

Software Engineer, smart information services GmbH, Potsdam, Germany 1997

Student Assistant, Northwestern University Library 1995 – 1996

EDUCATION

Dr. phil., Linguistics (summa cum laude), Universität Potsdam 2004 – 2010

- Dissertation: *Finite-State Canonicalization Techniques for Historical German*
  Advisor: Prof. Dr. Peter Staudacher

Diplom, *Computational Linguistics*, Universität Potsdam 1996 – 2002

- *Diplom* Thesis (2001), *Relational Query Feature Structures*

B. A., Philosophy; minor, Cognitive Science, Northwestern University 1992 – 1996

HONORS AND AWARDS

Daniel Bonbright Scholar, awarded 1996 by Northwestern University, College of Arts and Sciences.

Phi Beta Kappa, elected 1996 by Phi Beta Kappa chapter Alpha of Illinois, Northwestern University.

URLS

- **Home**: `https://kaskade.dwds.de/~jurish/`
- **CPAN**: `https://metacpan.org/author/MOOCOW`
- **sourceforge**: `https://sourceforge.net/u/mukau/`
- **github**: `https://github.com/moocow-the-bovine`

TEACHING

7th European Summer University in Digital Humanities, "Culture & Technology",
University of Leipzig, Leipzig, Germany:

- 2016 July *Searching Linguistic Patterns in Large Text Corpora for Digital Humanities Research*; (together with Erhard Hinrichs, Lothar Lemnitzer, and Alexander Geyken)

Universität Potsdam, Potsdam, Germany:

- 2005 Spring *Grammar Induction*
- 2004 Fall *Practical Perl Programming*
- 2004 Spring *Statistical Methods in Computational Linguistics*
- 2003 Fall *Text-to-Speech Synthesis*
- 2003 Spring *Practical Perl Programming*
- 2001 Fall PROLOG *for Linguists* (Teaching Assistant / lab session)

STUDENTS & SUPERVISION

Master's students:

- Shirley Blau (Universität Potsdam)

Diplom students:

- Mustafa Simsek (Berlin Technical University / Universität Potsdam)

RESEARCH-RELATED ACTIVITIES

- 2020: Corpus infrastructure development and deployment for the "GEI-Digital" textbook corpus (`http://diacollo.gei.de/`) in cooperation with the *Georg-Eckert-Institut (Leibniz-Institut für Internationale Schulbuchforschung, GEI)*, Braunschweig, Germany.
- 2019: Co-organizer of the CLARIN Workshop on Natural Language Processing for Historical Documents, Berlin, Germany (with Martin Wynne and Christian Thomas).
- 2018: Peer reviewer for *Language Resources and Evaluation (LREV)*.
- 2016–2017: Cooperation with the *Hessisches Landesamt für geschichtliche Landeskunde (HLGL)* and the *Universtät Marburg* for orthographic normalization and robust automated linguistic annotation of historical archive data, with a focus on regional toponyms.
- 2016: Co-organizer of the SIGFSM Workshop on Statistical NLP and Weighted Automata (StatFSM 2016) at the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany (with Andreas Maletti, Uwe Springmann, and Kay-Michael Würzner).
- 2015: Program Committee member for the *12th International Conference on Finite-State Methods and Natural Language Processing (FSMNLP 2015)*, Düsseldorf, Germany.
- 2011: Reviewer for the Workshop *Lexical Resources in Psycholinguistic Research*, Berlin.
- 2008: Reviewer for the *9th Conference on Natural Language Processing* (KONVENS), Berlin.
- 2005: Reviewer for the *Workshop on Computational Modelling of Lexical Acquisition*, Split, Croatia
- 1997: Contributed classificatory patterns and algorithms to the FACILE (LE-2440) Project ("Fast and Accurate Categorization of Information by Language Engineering"), a cooperation between Quinary (Milan), IRST (Trento), UMIST (Manchester), SEMA (Madrid), sis (Berlin), Caja Segovia (Spain), and Italrating (Milan).

RESEARCH INTERESTS

- **Diachronic Distributional Semantic Modelling** **BBAW**
  Berlin, Germany
  since 2009

  Conventional distributional semantic modelling can be understood as a formalization of J. R. Firth's well-known quip that "you shall know a word by the company it keeps." A word's meaning may however change over time, due to acquisition of new readings, loss of archaic uses, or new demands of the external world to be described. In my work with the *Deutsches Textarchiv*, DWDS, CLARIN-D, and ZDL projects at the Berlin-Brandenburgische Akademie der Wissenschaften, I implemented a set of diachronic distributional semantic models: models which incorporate not only the lexeme (word) as occurring in the text but also the date of each occurrence, thus enabling linguists, lexicographers, and historians to achieve a clearer picture of diachronic changes in the word's usage, in particular those related to semantic shift or discourse environment.

- **Robust Canonicalization of Historical and Non-standard Text** **BBAW**
  Berlin, Germany
  since 2005

  Historical text presents unique problems for conventional morphological analysis techniques based on a canonical orthographic form, due to the fact that orthographic conventions vary widely with both time period of a text's origin and with its author. My research has shown that combining type-level conflation techniques such as conservative transliteration, phonetic identity, and a robust finite-state rewrite cascade with a token-level disambiguator in the form of a dynamic Hidden Markov Model can provide a reliable and application-independent estimate of extant canonical cognates for historical input, thus minimizing the need for specialized lexical resources. The developed techniques are currently being used in the *Deutsches Textarchiv* project[1] at the Berlin-Brandenburgische Akademie der Wissenschaften to index and query a corpus of historical German text. More recent work has focused on adapting and applying the established techniques to other non-standard text varieties, in particular computer-mediated communication text.

- **Corpus Infrastructure and Information Retrieval** **BBAW**
  Berlin, Germany
  since 2001

  The goal of the DWDS (*Digitales Wörterbuch der deutschen Sprache* "Digital Dictionary of the German Language") project is the development of a lexicographic database on the basis of a large text corpus drawn chiefly from the 20th century. The DTA (*Deutches Textarchiv*, "German Text Archive") project extended that corpus by digitizing, annotating, and publishing an interdisciplinary corpus of historical German text from the 17th, 18th, and 19th centuries. In both cases, the raw corpus material is encoded in the Extensible Markup Language (XML) according to the Text Encoding Initiative (TEI) standards. In order for the corpus to provide useful data to end users, however, it must be efficiently searchable.

  During my work with the DWDS project at the Berlin-Brandenburgische Akademie der Wis-

[1] `https://www.deutschestextarchiv.de`

senschaften (BBAW), I designed a relational normal form for corpus data, and implemented robust software tools for document import and indexing, as well as administration and maintenance of the underlying relational database. Additionally, I designed and implemented query interface modules in Perl and Oracle PL/SQL which were used in the construction of a web-based corpus search engine. Later work on extracting a corpus of quotation evidence from the *Deutsches Wörterbuch* led to the development of the flexible relational indexing architecture and query language `Taxi`. In 2011, I assumed responsibility for maintaining and further developing DDC, the corpus indexing and retrieval software used by the DWDS and DTA projects, extending its flexibility by implementing protocols for external query term expansion, used e.g. to account for historical spelling variation phenomena, as well as to provide thesaurus-based semantic search functionality. Since 2014, I have been responsible for design, implementation, deployment, and maintenance of the flexible *D** corpus management infrastructure used to provide a unified framework for all DDC corpora currently in use at the BBAW (ca. 43G tokens in ca. 113M documents total). Since 2019, the *D** framework has also been used for the corpora of the ZDL project, including automated regular updates of the corpus indices to incorporate new text data. The core *D** infrastructure has been available as a docker image since 2017, and has been actively deployed in docker containers hosted by cooperating projects outside the BBAW since 2020.

- **Dependency Induction** **Universität Potsdam**
  Potsdam, Germany
  since 2004

  The overall goal of this project is the construction of a large-scale finite-state dependency lexicon on the basis of only raw text input. Natural language exhibits a number of phenomena which can be broadly assimilated under the general notion of "dependency". Of these phenomena, perhaps the most well-known is that which instantiates the theoretical notion of a head, its complements and specifiers. In particular, dependency lexica for verbal heads have proved useful in establishing restrictions on potential verb-argument structures — both syntactic ("valences") and semantic ("selectional restrictions") — for use in both parsing and generation systems.

  Restricting input to only raw text attempts to address the question of how much *a priori* knowledge of a target language — or indeed of linguistic structure in general — is in fact required in order to acquire a viable inventory of lexical data. Use of finite-state devices (automata and transducers) to represent the target language data effectively restricts the class of learnable languages to type-3 (regular) languages, which are known to be insufficient for the accurate characterization of natural language phenomena. Nonetheless, previous work has shown that finite state devices can indeed provide useful "shallow" analyses of a considerable subset of natural language data, and this at a very low computational cost.

- **Improvisational and Musical Languages** **Universität Potsdam**
  Potsdam, Germany
  since 2004

  The main focus of this project is the characterization of generic musical structure in terms of the apparatus of formal language theory. Preliminary results suggest that musical structure falls into the same class as natural language with respect to strong generative capacity – the class of *mildly context-sensitive languages* described by Joshi.

An important outgrowth of this research is the precise characterization of the *improvisation property*, closely related to the *valid prefix property*. Roughly formulated, the improvisation property holds for all languages whose strings can be incrementally generated from left to right with linear time complexity. Psycholinguistic evidence suggests that natural languages do indeed display such a property, which is sometimes referred to as *online production*.

Future work will involve the formal characterization of as well as the development of generation and processing algorithms for classes of formal languages displaying the improvisation property in the general case, and application of these results to both natural and musical languages. Of particular interest in this respect is the intersection of the improvisational with the mildly context sensitive languages.

- **Morphologically Informed PoS-Tagging** **BBAW**
  Berlin, Germany
  since 2002

  The goal of this project is the development and implementation of improved techniques for Part-of-Speech (PoS) Tagging in the presence of a strong morphological analysis component, by integrating linguistically motivated rule-based approaches and stochastic modelling techniques.

  It was shown that extending a traditional Hidden Markov Model (HMM) tagger by incorporating *ambiguity classes* reduced tagging errors on German newspaper text by 17.6% with respect to a traditional, purely lexical HMM tagger. Additionally, use of the morphological component as a source of *a priori* information concerning the possible analyses for a given input token provides the extended tagger with a linguistically motivated means for search space reduction, enabling an improvement of up to 94.4% in computation speed.

  Work on this project led to the development of the `dwdst` / `moot` / `mootm` PoS tagging suite, which uses finite state technology to represent the morphological analysis component, and which is currently in use at the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW) for annotating large corpora of German text.[2,3]

  Future work will investigate the potential for further improvement by integration of a morphological component capable of returning partial results, providing some analysis even for unknown tokens, and thus enabling a potentially drastic reduction of system memory requirements.

- **Realtime Streaming Text-to-Speech Synthesis** **Universität Potsdam**
  Potsdam, Germany
  since 2002

  The goal of this project is the development of a text-to-speech (TTS) system capable of operating in a hard-realtime environment. While most currently available TTS systems operate chunk-wise on buffered input text (e.g. sentences), human speakers are fully able to pronounce visually presented text with natural intonation as the text is presented.

  Work on this project led to the development of `ratts`, the "Realtime Analogue Text-To-Speech" external library for Miller Puckette's realtime signal processing environment, Pd. Experience with this system has shown that the adaptation of existing TTS techniques to a

[2] `https://www.dwds.de`
[3] `https://www.deutschestextarchiv.de`

hard-realtime environment is indeed non-trivial, involving extensive structural and algorithmic modifications in order to accommodate the incomplete and unpredictable nature of the streaming text input model. Future research will extend these insights to other TTS techniques, making use of finite state technology to represent persistent processing data such as working memory and computation state.

- **Relational Databases and Linguistic Objects** **Universität Potsdam**
  Potsdam, Germany
  since 1997

  This goal of this ongoing project is to establish a clearly defined interface between linguistic objects and relationally structured data. This requires a formal definition and implementation of a relational semantics for linguistically motivated objects one the one hand, and the characterization of linguistic structures in terms of efficiently indexed relational data on the other hand.

  Work on this project led to the development of the relational database query generators UGH and QuD, which translate from feature structures (untyped and typed, respectively) to well-formed relational database queries expressed in the popular query language SQL. The high level of abstraction provided by the use of underspecified feature structures to represent database queries has proved useful in the construction of user interfaces to lexical databases such as CELEX, as well as in the construction of a full-fledged information retrieval system. Future research in this area will focus on efficient relational (index) structures for more traditional linguistic objects such as analysis feature structures, finite state machines, dependency graphs, and phrase-structure trees.

SKILLS

COMPUTER AND INFORMATION TECHNOLOGY:

- Expert knowledge of (weighted) finite-state natural language processing techniques and tools.
- Advanced knowledge of statistical language modeling techniques.
- Advanced knowledge of machine learning techniques, including data clustering, classification, distributional semantics, and collocation extraction.
- Advanced knowledge of text corpus formats, annotation, and processing techniques.
- Advanced knowledge of scalable information retrieval architectures, including relational databases, key-value stores, full-text indices, and native file formats.
- Operating systems: GNU/Linux, UNIX, MacOS, MS Windows, MS-DOS
  - Advanced administrative knowledge of GNU/Linux (esp. Debian, ubuntu)
- Advanced knowledge of software version control systems: CVS, SVN, git.
- Advanced knowledge of Interprocess Communications (IPC) techniques, including thread message queues, shared memory, pipes, FIFOs, sockets, and server-client architectures.
- Advanced knowledge of Common Gateway Interface (CGI) programming, tools, and techniques, including FastCGI and RESTful service architectures.
- Good knowledge of multi-threaded parallel program optimization techniques.
- Good knowledge of docker-based software packaging and deployment.
- Good knowledge of infrastructure monitoring software icinga.

- Good knowledge of jenkins automation framework.
- Advanced knowledge of the professional typesetting system LaTeX.
- Advanced knowledge of audio recording, processing, and analysis software; in particular real-time signal processing environments (Pd, MAX/MSP).
- Good knowledge of TCP/IP network architecture, administration, and maintenance.
- Good knowledge of conventional WYSIWYG office software packages.
- Good knowledge of markup and interchange formats: HTML, XML, TEI, JSON, YAML, *etc.*
  - Advanced knowledge of XML processing techniques with XSLT and Perl.
- Strong technical and scientific writing skills.

PROGRAMMING LANGUAGES:

- C, C++, Perl, XS, PDL, Python, SQL, JavaScript, jQuery, PHP, LISP, PROLOG, Java, and others.

NATURAL LANGUAGES:

- English (native, GER C2)
- German (accent-free, GER C1)

PUBLICATIONS

REFEREED PUBLICATIONS

Bryan Jurish. "Diachronic Collocations, Genre, and DiaCollo." In R. J. Whitt (editor), *Diachronic Corpora, Genre, and Language Change*, volume 85 of *Studies in Corpus Linguistics*, pages 42–64. Amsterdam, John Benjamins, 2018.
`https://dx.doi.org/10.1075/scl.85.03jur`

Alexander Geyken, Matthias Boenig, Susanne Haaf, **Bryan Jurish**, Christian Thomas & Frank Wiegand. "Das Deutsche Textarchiv als Forschungsplattform für historische Daten in CLARIN." In H. Lobin, R. Schneider, and A. Witt (editors), *Digitale Infrastrukturen für die germanistische Forschung*, volume 6 of *Germanistische Sprachwissenschaft um 2020*, pages 219–248. Berlin/Boston, De Gruyter, 2018.
`https://dx.doi.org/10.1515/9783110538663-011`

Bryan Jurish & Henriette Ast. "Using an Alignment-based Lexicon for Canonicalization of Historical Text." In J. Gippert and R. Gehrke (editors), *Historical Corpora: Challenges and Perspectives*, volume 5 of *Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)*, pages 197–208. Narr, Tübingen, 2015.
`https://cudmuncher.de/~moocow/pubs/ja2012using.pdf` (draft)

Bryan Jurish, Marko Drotschmann & Henriette Ast. "Constructing a canonicalized corpus of historical German by text alignment." In P. Bennett, M. Durrell, S. Scheible, and R. J. Whitt (editors), *New Methods in Historical Corpora*, volume 3 of *Corpus Linguistics and Interdisciplinary Perspectives on Language (CLIP)*, pages 221–234. Narr, Tübingen, 2013.
`https://cudmuncher.de/~moocow/pubs/jda2013constructing_draft.pdf` (draft)

Bryan Jurish. "Efficient online k-best lookup in weighted finite-state cascades." In T. Hanneforth and G. Fanselow, editors, *Language and Logos: Studies in Theoretical and Computational Linguistics*, volume 72 of *Studia grammatica*, pages 313–327. Akademie Verlag, Berlin, 2010.
`https://cudmuncher.de/~moocow/pubs/jurish2010kbest_draft.pdf` (draft)

REFEREED CONFERENCE PROCEEDINGS

Bryan Jurish and Maret Nieländer. "Using DiaCollo for historical research." *CLARIN Annual Conference 2019*, Leipzig, Germany, 30th September-2nd October, 2019.
`https://cudmuncher.de/~moocow/pubs/jn2019using.pdf`

Bryan Jurish. "Some remarks on text data visualization and codec transparency." *Visualisierungsprozesse in den Humanities: Linguistische Perspektiven auf Prägungen, Praktiken, Positionen (VisuHu 2017)*, Zürich, 17th-19th July, 2017.
`https://cudmuncher.de/~moocow/pubs/jurish-visihu2017-abstract.pdf`

Bryan Jurish. "Diachronic Collocations and Genre: a case for DiaCollo?" *Diachronic Corpora, Genre, and Language Change* Nottingham, UK, 8th-9th April, 2016.
`https://cudmuncher.de/~moocow/pubs/jurish2016genre.pdf` (draft)

Bryan Jurish, Alexander Geyken & Thomas Werneke. "DiaCollo: diachronen Kollokationen auf der Spur." *DHd 2016: Modellierung – Vernetzung – Visualisierung*, Leipzig, Germany 7th-12th March, 2016.
`https://cudmuncher.de/~moocow/pubs/jgw2016diacollo.pdf` (revised and corrected draft)

Bryan Jurish, "DiaCollo: On the trail of diachronic collocations'.' *CLARIN Annual Conference 2015*, Wrocław, Poland, 15th-17th October, 2015.
`https://cudmuncher.de/~moocow/pubs/jurish2015diacollo-clarin.pdf`

Kay-Michael Würzner & **Bryan Jurish**. "Dsolve – Morphological Segmentation for German using Conditional Random Fields." *Fourth International Workshop on Systems and Frameworks for Computational Morphology (SFCM)*, Stuttgart, Germany, 17th–18th September, 2015 (to appear).

Kay-Michael Würzner & **Bryan Jurish**. "A hybrid approach to grapheme-phoneme conversion." *12th International Conference on Finite State Methods and Natural Language Processing (FSMNLP)*, Düsseldorf, Germany, 22th–24th June, 2015.
`https://cudmuncher.de/~moocow/pubs/wj2015gramophone.pdf`

Bryan Jurish, Christian Thomas & Frank Wiegand. "Querying the deutsches Textarchiv." *Beyond Single-Shot Text Queries: Bridging the Gap(s) Between Research Communities (MindTheGap'14)*, Workshop held in conjunction with iConference'14, Berlin, Germany, 4th March, 2014.
`https://ceur-ws.org/Vol-1131/mindthegap14_7.pdf`

Bryan Jurish & Kay-Michael Würzner. "Multi-threaded composition of finite-state transducers." *11th International Conference on Finite State Methods and Natural Language Processing (FSMNLP)*, St Andrews, Scotland, 15th–17th July, 2013.
`https://www.aclweb.org/anthology/W13-1813`

Bryan Jurish. "Canonicalizing the deutsches Textarchiv." In I. Hafemann (editor), Proceedings of the conference *Perspektiven einer corpusbasierten historischen Linguistik und Philologie*, Berlin, Germany, 12th–13th December 2011; Volume 4 of *Thesaurus Linguae Aegyptiae*, Berlin-Brandenburgische Akademie der Wissenschaften, 2013.
`https://edoc.bbaw.de/volltexte/2013/2443/pdf/Jurish.pdf`

Alexander Geyken, Susanne Haaf, **Bryan Jurish**, Matthias Schulz, Jakob Steinmann, Christian Thomas, & Frank Wiegand. "Das Deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv." In S. Schomburg, C. Leggewie, H. Lobin & C. Puschmann (editors), Proceedings of *Digitale Wissenschaft: Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pages 157–161,

20th–21st September 2010; 2nd, expanded edition, 2011.
https://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf

Bryan Jurish. "Comparing canonicalizations of historical German text." In *11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON)*, pages 72–77, Uppsala, Sweden, 15th July, 2010.
https://www.aclweb.org/anthology/W10-2209

Bryan Jurish. "Finding canonical forms for historical German text" In A. Storrer, A. Geyken, A. Siebert, and K.-M. Würzner (editors), *Text Resources and Lexical Knowledge: selected papers from the 9th Conference on Natural Language Processing, KONVENS*, Berlin, Mouton de Gruyter, 2008.
https://cudmuncher.de/~moocow/pubs/jurish2008finding.pdf

Bryan Jurish. "Hybrid syntactic category induction." Paper presented at the *Workshop on Computational Modelling of Lexical Acquisition (CPALA)* Split, Croatia, July, 2005.
https://cudmuncher.de/~moocow/pubs/jurish2005hybrid_color.pdf

Bryan Jurish. "Music as a formal language." Paper presented at the *first international pd∼convention*, Graz, Austria, 27th September–3rd October, 2004. In F. Zimmer (editor), *bang | pure data*, Wolke Verlag, 2006.
https://cudmuncher.de/~moocow/pubs/pdconv04.pdf

Refereed Journal Publications

Alexander Geyken, Adrien Barbaresi, Jörg Didakowski, **Bryan Jurish**, Frank Wiegand & Lothar Lemnitzer. "Die Korpusplattform des 'Digitalen Wörterbuchs der deutschen Sprache' (DWDS)." *Zeitschrift für germanistische Linguistik*, 45(2):327–344, 2017.
https://dx.doi.org/10.1515/zgl-2017-0017

Bryan Jurish & Kay-Michael Würzner. "Word and Sentence Tokenization with Hidden Markov Models." *Journal for Language Technology and Computational Linguistics*, 28(2):61–83, 2013.
https://cudmuncher.de/~moocow/pubs/jw2013tokenization.pdf

Alexander Geyken, Susanne Haaf, **Bryan Jurish**, Matthias Schulz, Christian Thomas, & Frank Wiegand. "TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv." *Jahrbuch für Computerphilologie – online*, 2012.
http://computerphilologie.tu-darmstadt.de/jg09/geykenetal.html

Hans-Martin Gärtner & **Bryan Jurish**. "Postmodern linguistics and the prospects of neural syntax: Some polemical remarks." *Theoretical Linguistics* 37(1/2):37–44, 2011.

Bryan Jurish. "More than words: Using token context to improve canonicalization of historical German." *Journal for Language Technology and Computational Linguistics*, 25(1):23–40, 2010.
https://cudmuncher.de/~moocow/pubs/jurish2010more.pdf

Peter beim Graben, **Bryan Jurish**, Douglas Saddy & Stefan Frisch. "Language processing by dynamical systems." *International Journal of Bifurcation and Chaos* 14(2):599–621, 2004.
https://cudmuncher.de/~moocow/pubs/beimGrabenJurishEA2004.pdf

Invited Journal Publications

Bryan Jurish, "Tools, Toys, and Filters." *Rechtsgeschichte – Legal History*, Rg 24:347–348, 2016.
http://dx.doi.org/10.12946/rg24/347-348

THESES AND TECHNICAL REPORTS

Bryan Jurish. *Finite-state Canonicalization Techniques for Historical German.* Doctoral thesis, Universität Potsdam, 2012.
`http://opus.kobv.de/ubp/volltexte/2012/5578/`

Bryan Jurish. *A hybrid approach to part-of-speech tagging.* Final report, Project "Kollokationen im Wörterbuch", Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, 2003.
`https://cudmuncher.de/~moocow/pubs/dwdst-report.pdf`

Bryan Jurish. *Relational query feature structures*. Diplom thesis, Universität Potsdam, Institut für Linguistik, 2001.
`https://cudmuncher.de/~moocow/pubs/diplom/`

EDITED VOLUMES

Bryan Jurish, Andreas Maletti, Uwe Springmann, and Kay-Michael Würzner (eds). *Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata (StatFSM 2016)*. Workshop held at the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 12th August, 2016.
`http://www.aclweb.org/anthology/W/W16/W16-2400.pdf`

PRESENTATIONS

INVITED PRESENTATIONS

6th July, 2019: *Aktuelle Tendenzen der Diskurslinguistik: DiaCollo*. Julius-Maximilians Universität Würzburg, Germany.
`https://kaskade.dwds.de/~jurish/diacollo/diacollo-slides-2019-wuerzburg.pdf`

19th March, 2019: *DTA::CAB – a Field Spotter's Guide*. Universität Graz, Zentrum für Informationsmodellierung.
`https://cudmuncher.de/~moocow/software/dta-cab/workshop/`

19th–28th June, 2017: *Exploring diachronic collocations with DiaCollo*
`https://kaskade.dwds.de/~jurish/diacollo2017/`

- 19th June: Göttingen Centre for Digital Humanities, Universität Göttingen.
- 23rd June: Universität Potsdam, Institut für Linguistik.
- 28th June: Freie Universität Berlin, Institut für Romanische Philologie.

28th June, 2016: *DiaCollo*. Centre for Corpus Research, Univeristy of Birmingham, UK.
`https://cudmuncher.de/~moocow/pubs/diacollo-birmingham-slides.pdf`

30th September, 2004: *Music as a formal language: finite state automata and Pd*. First international pd convention, Graz, Austria.
`https://cudmuncher.de/~moocow/pubs/pdconv04talk.pdf`

CONFERENCE PRESENTATIONS

2nd October, 2019; joint work with Martin Wynne: *Natural Language Processing for Historical Documents.* Poster presentation of eponymous workshop at the CLARIN Annual Conference 2019, Leipzig, Germany.
`https://cudmuncher.de/~moocow/pubs/wj2019nlphist.pdf`

1st October, 2019; joint work with Maret Nieländer: *Using DiaCollo for historical research.* CLARIN Annual Conference 2019, Leipzig, Germany.

https://cudmuncher.de/~moocow/pubs/jn2019using.pdf

9th December, 2017; joint work with Maret Nieländer and Thomas Werneke: *DiaCollo and 'die Grenzboten'.* Genealogies of Knowledge I: "Translating Political and Scientific Thought across Time and Space," University of Manchester, UK.
https://cudmuncher.de/~moocow/pubs/jnw2017grenzboten-slides.pdf

23rd September, 2017; joint work with Thomas Werneke: *Visualizing Semantic Change with DiaCollo.* 20th International Conference on Conceptual History: "Concepts in the World – Politics, Knowledge, and Time," University of Oslo, Norway.
https://cudmuncher.de/~moocow/pubs/jw2017diacollo.pdf

9th April, 2016: *Diachronic collocations and genre: a case for DiaCollo?* Conference "Diachronic Corpora, Genre, and Language Change," Nottingham, UK.
https://cudmuncher.de/~moocow/pubs/jurish2016genre-slides.pdf

11th March, 2016; joint work with Alexander Geyken and Thomas Werneke: *DiaCollo: diachronen Kollokationen auf der Spur.* DHd 2016: "Modellierung – Vernetzung – Visualisierung," Leipzig, Germany.
https://cudmuncher.de/~moocow/pubs/jgw2016diacollo-slides.pdf

19th February, 2016: *Visualisierung diachroner Kollokationen mit DiaCollo.* Workshop "Die geisteswissenschaftliche Perspektive: Welche Forschungsergebnisse lassen Digital Humanities erwarten?" Akademie der Wissenschaften und der Literatur, Mainz, Germany.
https://cudmuncher.de/~moocow/pubs/jurish2016diacollo-gwp-slides.pdf

8th February, 2016; joint work with Thomas Werneke: *Computergestützte Analyse von Kollokationen im diachronen Verlauf.* Workshop "Digitale Geschichtswissenschaft – neue Tools für neue Fragen?" of the CLARIN-D Working Groups "Neuere Geschichte" and "Zeitgeschichte", Berlin-Brandenburgische Akademie der Wissenschaften.
https://cudmuncher.de/~moocow/pubs/jw2016diacollo-dgw-slides.pdf

30th October, 2015; joint work with Alexander Geyken: *"Neue Entwicklungen und Wege bei der Erstellung, Erweiterung und Nutzung von Korpora am Zentrum Sprache."* KobRA workshop "Neue Wege in der Nutzung von Korpora – Data-Mining für die textorientierten Geisteswissenschaften," Berlin-Brandenburgische Akademie der Wissenschaften.
https://cudmuncher.de/~moocow/pubs/gj2015entwicklungen.pdf

16th October, 2015: *DiaCollo: On the trail of diachronic collocations.*, CLARIN Annual Conference 2015, Wrocław, Poland.
https://cudmuncher.de/~moocow/pubs/jurish2015diacollo-clarin-poster.pdf

14th September, 2015: *DiaCollo: ein interaktives Werkzeug zur Extraktion und Exploration diachroner Kollokationen.* Workshop "Historische Semantik und Semantic Web" of the AG "Elektronisches Publizieren," Union der deutschen Akademien der Wissenschaften, Heidelberg, Germany.
https://cudmuncher.de/~moocow/pubs/jurish2015diacollo-heidelberg-slides.pdf

18th November, 2014: *Semantics, similarity, and corpus search in the Deutsches Textarchiv.* 2nd DTA- & CLARIN-D Conference and Workshop "Textkorpora in Infrastrukturen für die Geistes- und Sozialwissenschaften," Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, Germany.
https://cudmuncher.de/~moocow/pubs/jurish2014semantics-talk.pdf

17th November, 2014; joint work with Susanne Haaf: *Die Vielfalt vereinen: Die CLARIN-Eingangsformate CMDI und TCF.* 2nd DTA- & CLARIN-D Conference and Workshop "Textkorpora in Infrastrukturen für die Geistes- und Sozialwissenschaften," Berlin-Brandenburgische Akademie der Wissenschaften, Berlin, Germany.

4th March, 2014; joint work with Christian Thomas and Frank Wiegand: *Querying the Deutsches Textarchiv.* Beyond Single-Shot Text Queries: Bridging the Gap(s) Between Research Communities (MindTheGap'14), workshop held in conjunction with iConference'14, Berlin, Germany.
`https://cudmuncher.de/~moocow/pubs/jtw2014querying-talk.pdf`

23rd September, 2013; joint work with Kay-Michael Würzner, Maria Ermakova, and Sophie Arana; presentation by K.-M. Würzner: *Canonicalization techniques for computer-mediated communication.* Verarbeitung und Annotation von Sprachdaten aus Genres internetbasierter Kommunikation, workshop held at the GSCL Conference 2013, Darmstadt, Germany.
`https://cudmuncher.de/~moocow/pubs/jwea2013_gscl_ibk.pdf`

17th July, 2013; joint work with Kay-Michael Würzner: *Multi-threaded composition of finite-state transducers.* 11th International Conference on Finite State Methods and Natural Language Processing (FSMNLP), St Andrews, Scotland.
`https://cudmuncher.de/~moocow/pubs/jw2013multi-slides.pdf`

15th March, 2013; joint work with Kay-Michael Würzner, Lothar Lemnitzer, & Alexander Geyken; presentation by K.-M. Würzner: *Linguistic annotation of computer-mediated communication, (not only) an explorative analysis.* 35. Jahrestagung der deutschen Gesellschaft für Sprachwissenschaft (DGfS). Potsdam, Germany.
`https://cudmuncher.de/~moocow/pubs/ibk_dgfs2013.pdf`

6th March, 2013; joint work with and presentation by Christian Thomas: *Named Entity Recognition (NER) im Deutschen Textarchiv – Computerlinguistisch gestützte Identifikation von Personen- und Ortsnamen in den Korpora des DTA.* Workshop "Mehr Personen – Mehr Daten – Mehr Repositorien," Berlin, Germany.
`https://kaskade.dwds.de/dtaq/files/DTAE-NER_vortrag-2013-03-06.pdf`

28th September, 2012; joint work with Kay-Michael Würzner, Alexander Geyken, & Lothar Lemnitzer; presentation by K.-M. Würzner: *Kollaborative Erstellung eines annotierten Korpus als Grundlage für die Anwendung statistischer Ansätze der automatischen Sprachverarbeitung auf internetbasierte Kommunikation* Webkorpora in Linguistik und Sprachforschung, Mannheim, Germany,
`http://hypermedia.ids-mannheim.de/gscl-ak/Folien_Geyken.pdf`

6th December, 2012; joint work with Henriette Ast: *Using an alignment-based lexicon for canonicalization of historical text.* Historical Corpora 2012, Goethe University, Frankfurt am Main, Germany.
`https://cudmuncher.de/~moocow/pubs/ja2012using-slides.pdf`

29th May, 2012; joint work with Kay-Michael Würzner: *Multi-threaded composition of finite-state transducers.* 6th International Workshop on Weighted Automata Theory and Applications (WATA 2012), Dresden, Germany.
`https://cudmuncher.de/~moocow/pubs/mtfsm_talk_wata2012.pdf`

13th December, 2011: *Finite-state canonicalization techniques for historical German.* Perspektiven einer corpusbasierten historischen Linguistik und Philologie, Berlin, Germany.

30th April, 2011; joint work with Marko Drotschmann & Henriette Ast: *Constructing a canonicalized*

corpus of historical German by text alignment. New Methods in Historical Corpora, Manchester, UK.
`https://cudmuncher.de/~moocow/pubs/jda2011constructing-slides.pdf`

11th October, 2010: *More than words: Orthographic standardization in the* Deutsches Textarchiv. Workshop "Das Deutsche Textarchiv: Vernetzung und Nachnutzung," Berlin, Germany.
`https://cudmuncher.de/~moocow/pubs/jurish2010dtaws.pdf`

15th July, 2010: *Comparing canonicalizations of historical German text.* 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, Uppsala, Sweden.
`https://cudmuncher.de/~moocow/pubs/jurish2010comparing-slides.pdf`

30th September, 2008: *Finding canonical forms for historical German text*. 9th Conference on Natural Language Processing (KONVENS), Berlin, Germany.
`https://cudmuncher.de/~moocow/pubs/jurish2008finding-slides.pdf`

26th July, 2005: *Hybrid syntactic category induction*. Workshop on Computational Modelling of Lexical Acquisition (CPALA), Split, Croatia.
`https://cudmuncher.de/~moocow/pubs/talk-split-slides.pdf`

28th April, 2004; joint work with Peter beim Graben: *Context-free parsing by dynamical systems*. Workshop on Mathematical Methods in Computational Linguistics, Universität Potsdam, Potsdam, Germany.

19th September, 2003: *Part-of-Speech tagging with finite-state morphology*. Collocations and Idioms: Linguistic, Computational, and Psycholinguistic Perspectives, Berlin, Germany.
`https://cudmuncher.de/~moocow/pubs/kollok2003.pdf`

SELECTED SOFTWARE

ORIGINAL SOFTWARE

- `DiaColloDB` *("Diachronic Collocation Database")*: A suite of tools for extraction of significant collocates from a diachronic text corpus using either efficient native index structures or a DDC search-engine to acquire underlying frequency data.
  `https://kaskade.dwds.de/dstar/dta/diacollo`
- `DTA::CAB` *("Cascaded Analysis Broker")*: A command-line and client/server suite for robust and reliable orthographic canonicalization of historical input text, in Perl.
  `http://www.deutschestextarchiv.de/public/cab/`
- `D* Corpus Management Tools`: A robust and flexible framework using GNU make to bootstrap fully annotated DDC search-engine indices and associated auxiliary databases from raw text corpora provided as (untokenized, un-annotated) TEI-XML. Includes a minimalistic HTML-form front-end for testing and evaluation as well as high-level RESTful APIs suitable for re-use and integration into external front-end (web) environments.
  `https://kaskade.dwds.de/dstar/doc/`
- `dta-tokwrap`: A suite of C utilities together with a high-level object-oriented Perl module for linguistically salient serialization of arbitrary XML documents encoded according to the Text Encoding Initiative (TEI) P5 Guidelines.
  `https://cudmuncher.de/~moocow/software/dta-tokwrap/`
- `GFSM` *("GFSM Finite State Manipulation" library)*: A C library for representation and manipulation of (weighted) finite-state machines, using GLIB for low-level data structures. Includes `GFSMXL`, an extension library for online k-best string lookup operations in weighted finite-state transducer cascades.
  `https://cudmuncher.de/~moocow/projects/gfsm`
- `moot` *("moot Tagger")*: A C++ library and program suite for highly accurate part-of-speech tagging in the presence of a strong morphological component, using ambiguity classes to improve performance for unknown words. Includes classes and programs for supervised training, tagging, model compilation, evaluation, and dynamic modelling.
  `https://cudmuncher.de/~moocow/projects/moot`
- `mootm` *("moot Tagger Morphology")*: A high-level C++ library and program suite for language independent morphological analysis using the GFSM C library for representation of the runtime language-specific morphological transducer.
  `https://cudmuncher.de/~moocow/projects/mootm`
- `unicruft`: C library and command-line utilities for fast conservative transliteration from UTF-8 to ASCII, Latin-1, or the subset of Latin-1 used in contemporary German orthography.
  `https://cudmuncher.de/~moocow/software/unicruft/`
- `WASTE` *("Word And Sentence Tokenization Estimator")*: A framework for detecting word and sentence boundaries in raw text using a Hidden Markov Model to estimate boundary placement in a stream of candidate word-like segments returned by a low-level rule-based scanner stage, developed in cooperation with Kay-Michael Würzner. Pre-built WASTE models exist for a number of languages, and additional models can be defined for various languages, genres, orthographic conventions, and/or target boundary-placement conventions with appropriate training material. WASTE is currently implemented as an extension to the `moot` part-of-speech tagging library.
  `https://www.dwds.de/waste/`
- `MUDL` *("MUDL Unsupervised Dependency Learning")*: A suite of Perl modules and executable scripts for unsupervised learning from raw text input. Includes among others classes for tok-

enization, n-gram modelling, smoothing, clustering, classification, and induction of finite-state automata.

- `DocClassify`: A suite of Perl modules and command-line utilities for fast runtime document classification and similarity-based k-nearest neighbor search using PDL for efficient representation of the underlying high-dimensional vector-space models. `DocClassify` has been successfully used in the *Deutsches Textarchiv* project for similarity-based automatic recommendation, and by a large German social media portal for user-specific advertising-relevant classification.
- `Taxi` *("Text and XML Index")*: Command-line and client/server suite for flexible indexing and intuitive query of structured linguistic data extracted from arbitrary XML documents using the mysql relational database as a back-end.
  `https://cudmuncher.de/~moocow/software/Taxi-Mysql/`
- `Lingua::LTS`: A Perl module providing an object-oriented compiler and interpreter for deterministic letter-to-sound rules with `festival`-like syntax and semantics using the `gfsm` library. A straightforward extension to the construction for precedence-ordered deterministic LTS rules as used by `festival` enables compilation of weighted parallel rule-sets into non-deterministic weighted finite state transducers.
  `https://cudmuncher.de/~moocow/software/Lingua-LTS/`
- `SayWhat`: A suite of Perl modules for context-free string generation and audio synthesis from a user-defined grammar. Includes both a graphical interface using perl/Tk and a scriptable command-line interpreter.
  `https://cudmuncher.de/~moocow/projects/saywhat`
- `ratts` *("Realtime Analogue Text-to-Speech")*: An external library written in C for Pd, Miller Puckette's realtime signal processing environment. Based on Nick Ing-Simmons' `rsynth` program. Includes realtime-safe, stream-capable Pd objects for tokenization, rule-based phonetization, and formant speech synthesis.
  `https://cudmuncher.de/~moocow/projects/pd`
- `QuD` *("Query Description" library)*: A library of Perl modules for translating between a representation of relational database queries as (underspecified, typed) feature structures and well-formed SQL queries. Includes abstract classes for representation and manipulation of bounded complete partial orders (type hierarchies) and (typed) feature structures.
  `https://cudmuncher.de/~moocow/pubs/diplom/modules`

MAINTAINED SOFTWARE

- `DDC` *v2.x*: An efficient and scalable corpus indexing and retrieval engine originally written by Alexey Sokirko. The 2.x branch of DDC supports multiple quasi-independent token-level attributes, flexible HTTP-based online query-term expansion, transparent content caching via `mmap(2)`, as well as index fragmentation to take advantage of contemporary multi-threaded server hardware and distributed corpora.
  `https://cudmuncher.de/~moocow/software/ddc`
- `gramophone`: An open-source package for hybrid grapheme-to-phoneme conversion using a set of heuristic mappings to determine admissible segmentations, a Conditional Random Field model for labeling candidate segmentations, and a language model over (*grapheme*,*phoneme*) segment-pairs to determine the optimal transcription. Developed in cooperation with Kay-Michael Würzner, `gramophone` is implemented using wapiti, OpenFst, OpenGrm, Python, and Perl.
  `https://kaskade.dwds.de/gramophone/`

Dr. phil. Bryan Jurish

Amselhainstraße 86 · 14612 Falkensee · Germany moocow@cudmuncher.de

REFERENCES

Johannes Bubenzer
2txt GmbH
bubenzer@2txt.de
Managing Director

Kay-Michael Würzner
Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden
kay-michael.wuerzner@slub-dresden.de
Department Head "Open Science", Specialist for German Linguistics and Literature

Prof. *em.* Dr. Peter Staudacher
Universität Potsdam, Germany
staudach@uni-potsdam.de
Professor (emeritus) of Computational Linguistics (Formal Languages and Automata)

Prof. Dr. Hans-Martin Gärtner
Hungarian Academy of Sciences
gaertner@nytud.hu
Research Professor of Formal Grammar and Pragmatics

Prof. Dr. Marcus Kracht
Universität Bielefeld
marcus.kracht@uni-bielefeld.de
Professor of Computational Linguistics and Mathematical Linguistics

Prof. Dr. Christiane Fellbaum
Princeton University
fellbaum@princeton.edu
Senior Research Scholar