38. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft - Sektion CL, Konstanz 24.-26. Februar 2016

A workflow for creating, analysing, and storing multi-layer corpora: Pepper, Atomic, ANNIS and LAUDATIO

Stephan Druskat[1,2], Thomas Krause[1], Carolin Odebrecht[1], Florian Zipser[1]

[1]Humboldt-Universität zu Berlin, [2]Friedrich-Schiller-Universität Jena

A workflow for creating, analysing, and storing multi-layer corpora

- The creation and analysis of corpus linguistic resources pose technical challenges: Different tools and formats have to be combined in a single workflow
- These challenges can best be faced with a generic architecture and metamodel, common to the complete tool chain
- We present an open source set of tools which support the conversion (**Pepper**, Zipser et al. (2011)), annotation (**Atomic**, Druskat et al. (2014)), analysis (**ANNIS**, Krause & Zeldes (2014)), and long-term accessibility (**LAUDATIO**, Odebrecht et al. (2015)) of corpora
- Our tools are well-aligned due to a common, generic graph-based data model, **Salt** (Zipser & Romary, 2010), which is theory-neutral and supports annotation types which can be represented as key-value pairs
- Our tools can be freely combined to represent a complete, iterative workflow for the creation of corpus linguistic resources (cf. central graphic)

Pepper

Pepper is a Swiss army knife for the conversion of corpora from one linguistic format into another. It comes to the rescue whenever your annotation tool produces a format your analysis tool cannot read.

- Converts corpora from and into **many linguistic formats**: Elan, CoNLL, MMAX2, ANNIS, Gate, RST, TCF, CoraXML, TreeTagger, Aldt, UAM, EXMARaLDA, generic XML, PTB, PAULA, TEI (subset), Tiger XML, txt, SaltXML, …
- **Merges corpora** from different formats into a single multi-layer corpus via its merging module
- Assists in the **documentation of your corpus** via its info module (Voigt et al., 2016)
- Uses Salt as a common meta model and is therefore **independent of tagsets or linguistic theories**
- **Extensible** for further formats and tasks via plug-in mechanism
- Available as **stand-alone** tool
- Can be integrated into your own tool as a **Java library**
- **Open source** and free (Apache License, source code on GitHub)
- Runs on **all major OSs** (written in Java)

Try it now!

- Download: **corpus-tools.org/pepper**
- Join the project and contribute

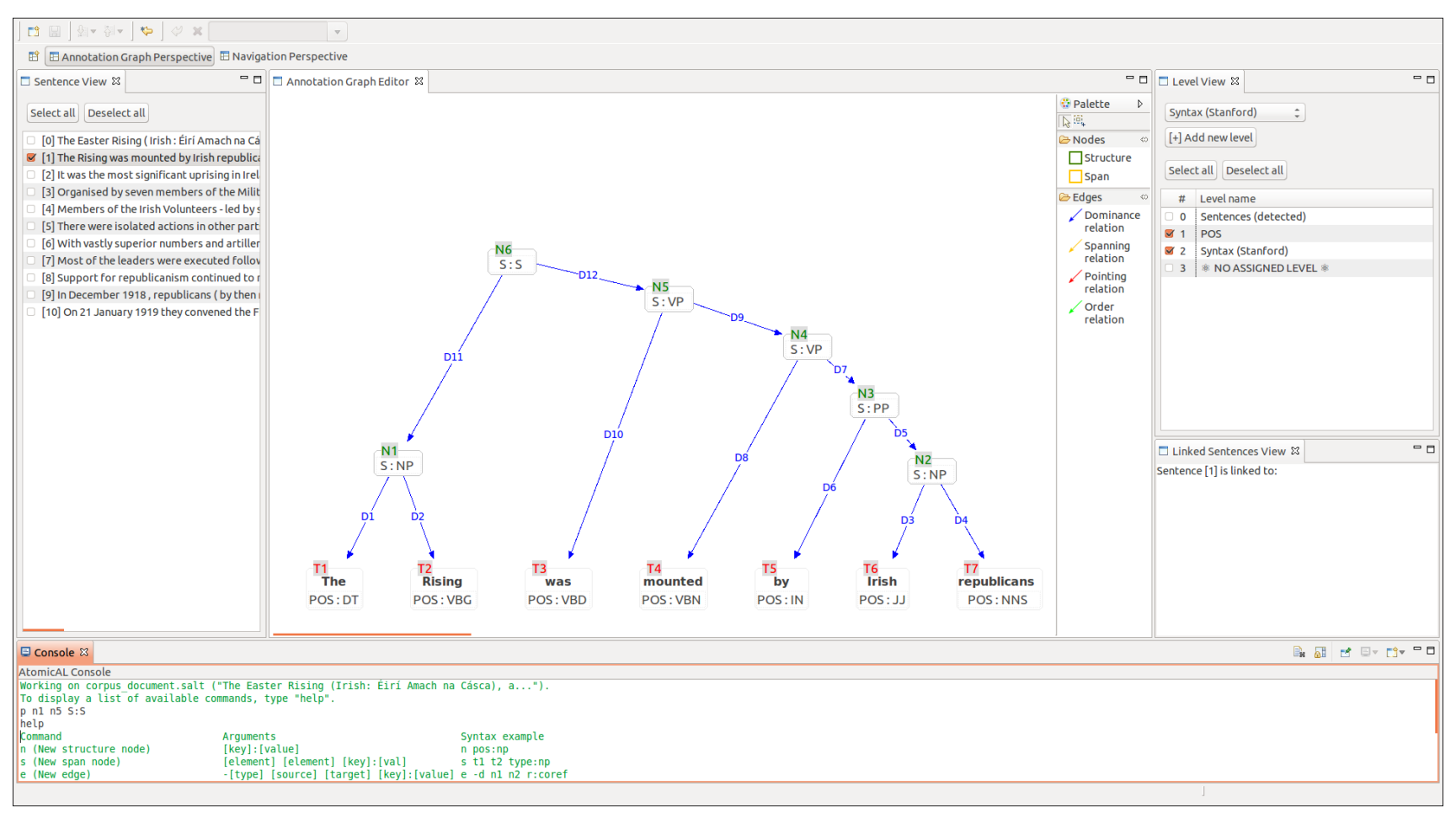
Atomic

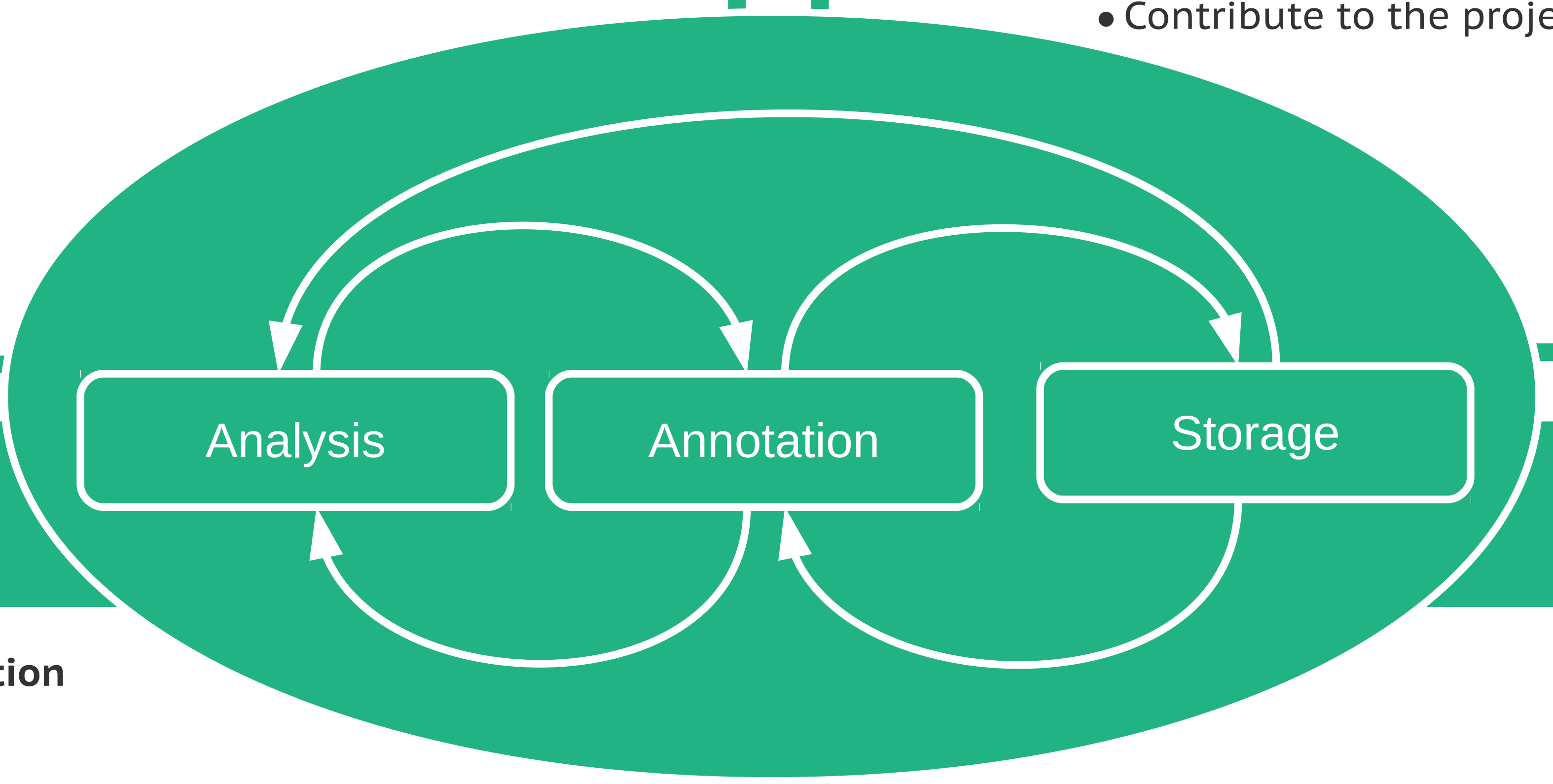
Atomic is a cross-platform multi-layer corpus annotation tool – and extensible platform – for the desktop.

- **True multi-layer support** due to its generic data model (Salt)
- **Theory-neutral annotation** independent of tagsets & annotation types
- Includes Pepper – controlled via a GUI – and is thus **compatible** with many linguistic formats
- **Annotation graph editor**, controllable via mouse and interactive command line
- Highly **extensible** due to its plug-in architecture (Eclipse RCP)
- Many **plug-ins available** (e.g., XML editors, SCM, real-time collaboration)
- **Experimental release** available, stable 1.0 version later in 2016
- **Open source** and free (Apache License, source code on GitHub)
- Runs on **all major OSs** (written in Java)

Try it now!

- Download: **corpus-tools.org/atomic**
- Contribute to the project





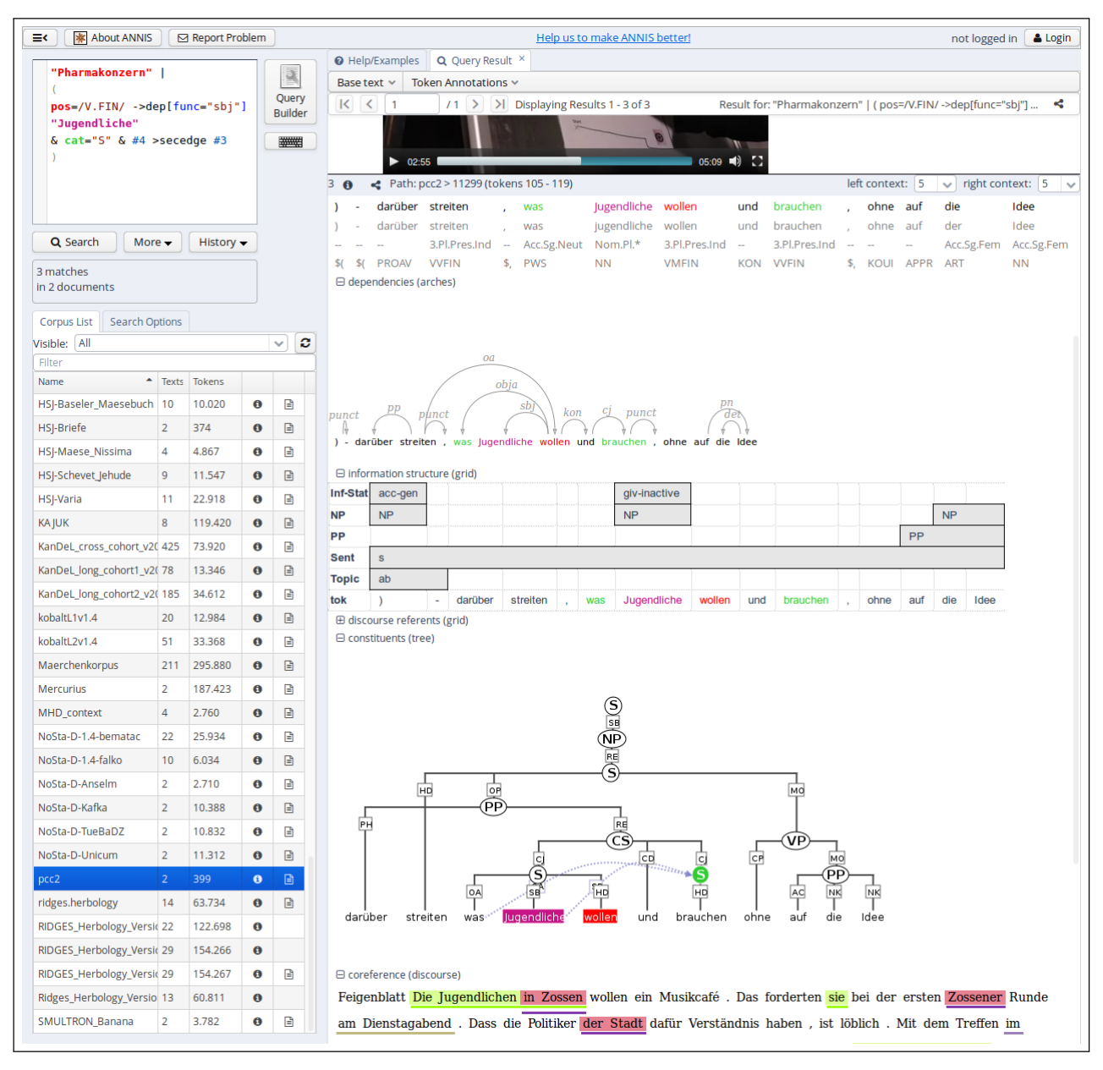
ANNIS

ANNIS is a web browser-based search and visualization tool for complex multi-layer linguistic corpora with diverse types of annotation.

- Addresses the need to **query and visualize** data from such varied areas as syntax, semantics, morphology, prosody, referentiality, lexis and more
- Support for **text** as well as **audio and video** corpora
- Generic, **native query language** which supports, e.g., tokens, spans, word forms, annotations, trees, pointing relations, regex searches, …
- Queries can be entered directly or constructed with an **intuitive graphical query builder**
- **Visualizations** for different annotation types: KWIC, trees, grids, grid trees, discourse views, dependencies, graphs; further visualizations can be added as plug-ins
- **Frequency analysis** using corpus metadata or annotations
- **Uses Salt** as interchange format between front- and backend, and for different visualizations
- Installable **locally or on a server**
- **Open source** and free (Apache License, source code on GitHub)
- Runs on **all major OSs** (written in Java)

Try it now!

- Download: **corpus-tools.org/annis**
- Join the project and contribute



LAUDATIO

The LAUDATIO repository is an open access, interdisciplinary research data repository for historical data.

- Allows for **interdisciplinary exchange** of data from historical linguistics
- Aims at a **unified description for corpora** from diverse disciplines and their different registers, languages, annotations, and formats
- Takes into account the **requirements of historical corpora**: non-standard variety data, transcription levels, etc.
- Extensive, technically abstract **metadata schema**, capturing the data's complete lifecycle, and accounting for language, places, dates, text types, data preparation methods, tools, formats, annotation guidelines, bibliographic metadata, research context, etc.
- Flexible and appropriate **documentation schema** with a subset of TEI customized by TEI ODD in a RELAX NG schema
- Includes an **interface with ANNIS** for complex and comprehensive search
- Accepts corpora licensed under a **Creative Commons** license
- **Open source** and free (Apache License, source code on GitHub)

Try it now!

- Browse **laudatio-repository.org**
- Consider submitting your own corpus

References

Druskat, Stephan, Lennart Bierkandt, Volker Gast, Christoph Rzymski & Florian Zipser. 2014. Atomic: an open-source software platform for multi-level corpus annotation. In Josef Ruppert & Gertrud Faaß (eds.), Proceedings of the 12th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2014), 228–234.

Krause, Thomas & Amir Zeldes. 2014. ANNIS3: A new architecture for generic corpus query and visualization. Digital Scholarship in the Humanities. `http://dx.doi.org/10.1093/llc/fqu057`.

Odebrecht, Carolin, Thomas Krause & Anke Lüdeling. 2015. Austausch von historischen Texten verschiedener Sprachen über das LAUDATIO-Repository. Poster presented at 37. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, 5 March, Leipzig University, Leipzig, Germany.

Voigt, Vivian, Florian Zipser & Carolin Odebrecht. 2016. SaltInfoModule - the x-ray to your corpus. Poster presented at 38. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, 25 February, Konstanz University, Konstanz, Germany.

Zipser, Florian & Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In Proceedings of the Workshop on Language Resource and Language Technology Standards, Seventh International Conference on Language Resources and Evaluation (LREC 2010), Valletta, Malta.

Zipser, Florian, Amir Zeldes, Julia Ritz, Laurent Romary & Ulf Leser. 2011. Pepper: Handling a multiverse of formats. Poster presented at 33. Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft, 24 February, Göttingen University, Göttingen, Germany.

http://corpus-tools.org

DFG

SFB 632 "Information Structure"

http://laudatio-repository.org