UNIVERSITÄT
ZU KÖLN

# Wissenschaftliche Literatur

## HS Anwendungen der Computerlinguistik

Nils Reiter
nils.reiter@uni-koeln.de

24. Oktober 2024

INSTITUT FÜR
DIGITAL HUMANITIES
UNIVERSITÄT ZU KÖLN

Was ist wissenschaftliche Literatur?
Was würden Sie Erstsemester:innen zum Umgang mit Literatur raten?
Was hätten Sie gerne früher gewusst?

# Section 1

## Overview

# Scientific Literature

Two core requirements

1. Quality assurance – reviewing
2. Long-term availability – archiving

# Peer Review

- Scientific articles are reviewed by other researchers/scientists
- Blindness
  - Double blind: Reviewer and authors are anonymous
  - Single blind: Only reviewers are anonymous
  - Zero blind / „Open Review“: No one is anonymous
- Different fields have different preferences
  - and different people have different preferences
  - CL: Double-blind (recently reaffirmed)
    - But: Preprint servers are an important venue in machine learning!

# Publication Venues

- Monographs (books): Except for theses, typically not reviewed
- Journal articles: Peer reviewed (details are journal-dependent)
- Conference articles: Peer reviewed (details are conference-dependent)
  - „Proceedings“ = Collection of all conference articles

# Publication Venues

- Monographs (books): Except for theses, typically not reviewed
- Journal articles: Peer reviewed (details are journal-dependent)
- Conference articles: Peer reviewed (details are conference-dependent)
  - „Proceedings“ = Collection of all conference articles

## Lengths and „Abstracts“

- Length varies
  - Conference articles $< 10$ pages
  - Journal articles ca. $10 - 50$ pages
- „Abstract“
  - Literal meaning: A summary of an article
  - Conference abstracts (DHd/DH) $\simeq$ short articles

# Relevant Publication Venues for CL

- Conferences
  - ACL / NAACL / EACL / EMNLP: Conferences (double-blind)
    - Association for Computational Linguistics
    - ⚠ ACL 2022: 604 long papers – ACL 2002: 65 papers

aclanthology.org

# Relevant Publication Venues for CL

- Conferences
  - ACL / NAACL / EACL / EMNLP: Conferences (double-blind)
    - Association for Computational Linguistics
    - ⚠ ACL 2022: 604 long papers – ACL 2002: 65 papers
    - Co-located workshops with more specific focus
    - „Workshop“ in CL: Mini conference
    - Workshops associated with *CL conferences also in anthology
  - COLING, KONVENS, LREC: Smaller conferences

aclanthology.org

# Relevant Publication Venues for CL

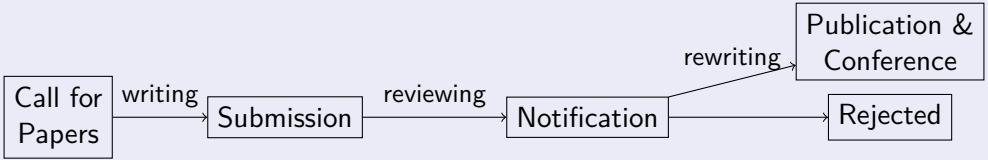
- Conferences
  - ACL / NAACL / EACL / EMNLP: Conferences (double-blind) aclanthology.org
    - Association for Computational Linguistics
    - ⚠ ACL 2022: 604 long papers – ACL 2002: 65 papers
    - Co-located workshops with more specific focus
    - „Workshop“ in CL: Mini conference
    - Workshops associated with *CL conferences also in anthology
  - COLING, KONVENS, LREC: Smaller conferences
- Journals: Uncommon
  - Computational Linguistics direct.mit.edu/coli
    - Also in anthology: `https://aclanthology.org/venues/cl/`
    - Fully open access

# Relevant Publication Venues for DH

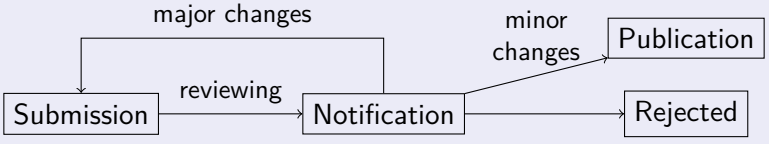
- New field, not yet fully established
- Venues from original H discipline (e.g., Journal of Literary Theory)
- DH, DHd
- Digital Scholarship in the Humanities (Literary and Linguistic Computing) academic.oup.com/dsh
  - Partially open access via UB
- Journal of Computational Literary Studies jcls.io
- DFG (funding agency): No reviewing → no worth
- Blogs – it depends on their authors
- Sammelbände / collections

# Publication Process and Timeline

## Conferences




## Journals



# Preprint-Servers

- Origin: Share preprints freely
- No review: Everyone can upload anything
- Popular for machine learning innovations
- Many papers are later/also submitted to a conference

arxiv.org

# Non-Scientific Literature

- Categories
  - All media for the general public (including newspapers and special interest journals)
    - E.g., Die Zeit, Segeln, GEO, ...
  - Blogs, YouTube channels and social media postings
    - E.g., spreeblick.com
  - Companies, lobby groups
  - Government publications

# Non-Scientific Sources

- Is it ok to use non-scientific sources?
  - It depends
- When is it ok?
  - When we are explicitly dealing with public opinion or reception
  - When we are looking into other disciplines
  - When a scientific discourse on a topic does not exist
  - When the topic in question is not a scientific question
- What are better sources than others?
  - Sources with references
  - Sources with scientific references
  - Sources with (scientific) references that are correctly reproduced

# Original/Primary Sources

- Scientific papers are usually written by the people who came up with the content
- Non-scientific sources are often mediated
- Original sources are better sources
  - Mediations may misrepresent the content substantially
- Examples
  - Press releases by universities about studies their researchers conducted
  - News articles about laws or decrees
  - Summaries of interviews

# Finding Literature

- Specialised repositories
  - Computational Linguistics aclanthology.org
  - Digital Humanities DH Index
  - Generic preprints arxiv.org
- References of other papers
- Your library USB Köln
  - Don't underestimate the ebook collection!
- Search engines Google Scholar Semantic Scholar
  - ⚠ Google finds a lot of non-scientific literature
- Wikipedia pages have often very good references

Section 2

# Reading Scientific Literature

# How to Read?

- Reading literature is work
- A work environment is important
- Reading multiple times is often necessary

# How to Read?

- Reading literature is work
- A work environment is important
- Reading multiple times is often necessary

## References

- Scientific references consist in:
  - Markers in the text (e. g., „Doe (2015)“ oder „[3]“)
  - Bibliographic details at the end
- Different styles
  - CL/DH: author-year
- URLs or DOIs
  - https://www.example.com
  - 10.1515/9783110693973 ⇒ `https://doi.org/10.1515/9783110693973`

# Scientific References

Daniel Preoţiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). „Automatically Identifying Complaints in Social Media“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf

# Scientific References

Daniel Preoţiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). „Automatically Identifying Complaints in Social Media“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf

Axel Pichler/Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003

# Scientific References

Daniel Preoţiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). „Automatically Identifying Complaints in Social Media“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf

Axel Pichler/Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003

Bei Yu (2014). „Language and gender in Congressional speech“. In: *Literary and Linguistic Computing* 29.1. _eprint: http://llc.oxfordjournals.org/content/29/1/118.full.pdf+html, pp. 118–132. DOI: 10.1093/llc/fqs073. URL: http://llc.oxfordjournals.org/content/29/1/118.abstract

# Scientific References

Daniel Preoţiuc-Pietro/Mihaela Gaman/Nikolaos Aletras (2019). „Automatically Identifying Complaints in Social Media“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf

Axel Pichler/Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003

Bei Yu (2014). „Language and gender in Congressional speech“. In: *Literary and Linguistic Computing* 29.1. _eprint: http://llc.oxfordjournals.org/content/29/1/118.full.pdf+html, pp. 118–132. DOI: 10.1093/llc/fqs073. URL: http://llc.oxfordjournals.org/content/29/1/118.abstract

Andrew Piper (2018). *Enumerations. Data and Literary Study*. University of Chicago Press

# Guiding Questions for CL/technical DH Papers

You should be able to answer (at least) these questions

- What was the task/the problem to be solved?
- What is the new aspect compared to previous research?
- How well did it work?
  - ⚠ Authors have an interest to highlight success and neglect failure
- Which experiments were made to measure it?
  - Which data and evaluation metrics were used?

# Critical Reflection of Technical Literature

- Was there an easier way to achieve similar performance?
- How many assumptions are incorporated (maybe implicit)?
  - What would be needed to redo it from scratch?
  - What would be needed to adapt it to another language/genre/domain?
- Why did the authors did it the way they did?
- Can the experiments actually show what the authors claim they show?
- Are the experiments „correctly“ interpreted? Are there alternative interpretations that are just as reasonable?
- Is there evidence to generalize results to „the language“, „the text type X“, …?

# Reading Non-Scientific Literature

- Make a local copy, ideally with all the meta data
  - No one guarantees that it is still there tomorrow
- Who wrote it and why?
- Do they know what they are writing about?
- What's the track record of the author/venue?
- Are author/venue involved in any way?

# References I


- Pichler, Axel/Nils Reiter (2020). „Reflektierte Textanalyse“. In: *Reflektierte Algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt*. Ed. by Nils Reiter/Axel Pichler/Jonas Kuhn. Berlin: De Gruyter, pp. 43–60. DOI: 10.1515/9783110693973-003.
- Piper, Andrew (2018). *Enumerations. Data and Literary Study*. University of Chicago Press.
- Preoţiuc-Pietro, Daniel/Mihaela Gaman/Nikolaos Aletras (2019). „Automatically Identifying Complaints in Social Media“. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 5008–5019. DOI: 10.18653/v1/P19-1495. URL: https://www.aclweb.org/anthology/P19-1495.pdf.

# References II


Yu, Bei (2014). „Language and gender in Congressional speech“. In: *Literary and Linguistic Computing* 29.1. _eprint: http://llc.oxfordjournals.org/content/29/1/118.full.pdf+html, pp. 118–132. DOI: 10.1093/llc/fqs073. URL: http://llc.oxfordjournals.org/content/29/1/118.abstract.