A Game with Complex Rules: Literature References in Literary Studies

Frederik Arnold[1], Robert Jäschke[1]

[1]Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

Abstract

Existing systems for reference extraction and segmentation are mostly tailored towards STEM fields (science, technology, engineering, medicine) and social sciences and can not properly handle references in literary studies. We present our annotation guidelines for literature references in literary studies and give an overview of difficult cases we encountered when creating a corpus of annotated scholarly works for literary studies. Specifically, we present challenges and requirements we identified for reference extraction and segmentation from scholarly articles in the field of literary studies.

Keywords

Literary studies, Reference extraction, Reference segmentation, Reference annotation

1. Introduction

The identification of communities of practice [1] in the field of modern German literary studies is an active area of research. Knowledge of subgroups of researchers that interpret similar texts, use similar methods, or focus on similar research questions would help to better understand the practice of interpretation. Communities of practice can be analyzed with *bibliometric* and *scientometric* approaches as Doerfel et al. [2] have done for the Formal Concept Analysis community and Jannidis et al. [3] for co-authorship in German literary studies. To apply such approaches at a large scale, methods to automatically extract literature references are necessary. Reference extraction and segmentation is the process of identifying and extracting references from text and then segmenting the identified references into individual components, such as, author, title, volume, etc.

Existing systems for reference extraction and segmentation are mostly tailored towards STEM fields (science, technology, engineering, medicine) (cf. Grobid [4], Prasad et al. [5]) or social sciences (cf. Hosseini et al. [6]). In the project *What matters? Key passages in literary works*,[1] we deal with references from scholarly works from German literary studies, specifically interpretation texts. These are not properly handled by existing approaches for they differ from other fields in some key aspects.

Existing systems work in two stages. The first stage is to analyze the layout of the PDF file

Understanding Literature References in Academic Full Text (ULITE)

frederik.arnold@hu-berlin.de (F. Arnold); robert.jaeschke@hu-berlin.de (R. Jäschke)

https://hu.berlin/fa (F. Arnold); https://hu.berlin/rj (R. Jäschke)

0000-0002-0417-4054 (F. Arnold); 0000-0003-3271-9653 (R. Jäschke)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://www.projekte.hu-berlin.de/en/schluesselstellen/index.html

and detect areas which are part of a literature reference, for example, a dedicated reference section or footnotes. The second step is to segment identified references. In this paper, we will ignore the first step, which in itself is difficult for scholarly works in literary studies and focus on the second step.

Our contributions are an analysis of literature references in literary studies and the identification of the challenges existing (or future) automatic approaches and annotators face. We provide selected examples that demonstrate those challenges. As a first step towards solving these issues, we developed annotation guidelines for literature references in literary studies alongside a first pilot annotation of eight fully and two partially annotated scholarly works. Finally, we want to spark a discussion on how to tackle automatic identification and extraction of such references.

This paper is organized as follows: In Section 2 we provide an overview on existing systems and corpora for literature reference detection and parsing. In Section 3, we describe our dataset and annotation approach. In Section 4, we summarize our findings on the challenges with examples and present our annotation guidelines. The paper concludes with a discussion and outlook on future work in Section 5.

2. Related Work

There are a number of datasets for reference extraction and segmentation. For example, UMass [7] is a dataset with 1 829 labeled reference strings from STEM fields. CORA[2] is a small dataset containing labeled reference strings from computer science. The Grobid dataset [4] is a combination of other datasets including CORA, UMass, and PubMed[3] (biomedical). The focus of these datasets is on reference strings from publications in English in STEM or related fields and they contain few to no footnotes. This makes them not suitable for our use case. One of the few German datasets[4] was created by Körner et al. [8], consisting of 125 German publications from the social sciences based on publications from the SSOAR repository.[5] Only 20 of the 125 publications contain references in footnotes. This is the most promising dataset for our use case. It is still limited as it is not diverse enough and internal references (see Section 4.2.2) are not annotated. Colavizza and Romanello [9] present a dataset of manually annotated references from books and journal articles on the history of Venice, mostly in Italian. This is one of the few datasets focused on articles from the humanities.

The fact that manual annotation is resource and time consuming has led to other approaches for creating datasets of annotated reference strings. Grennan et al. [10] introduce GIANT, a synthetic dataset of 991 411 100 labeled reference strings. These reference strings were automatically created from CrossRef[6] entries in combination with 11 564 citation styles. Thai et al. [11] introduce a similar approach and collected 6 023 publicly available BibTeX files from different websites which are then used to generate around 41 million labeled reference strings in 26 styles. These approaches can not easily be adapted to our domain. There are no styles available

[2] https://people.cs.umass.edu/~mccallum/data.html
[3] https://pubmed.ncbi.nlm.nih.gov/download/
[4] https://github.com/behnam2014/ssoar-gold-standard
[5] https://www.ssoar.info/
[6] https://www.crossref.org/

that could be used for automatic generation and references are more language-dependent, less structured and less generalizable.

As part of the EXCITE project,[7] Körner et al. [8] developed RefExt as a new approach to extract reference strings from German scholarly publications. The EXCITE project also developed the EXCITE toolchain [6]. For a technical background, see [12]. The EXCITE toolchain uses CERMINE [13] to extract text from PDF files. CERMINE is a Java library for extracting text from PDFs. CERMINE zone classification was trained on 2 500 documents from the GROTOAP2 dataset [14] which leads to CERMINE struggling with footnote extraction and in turn leads to difficulties for the complete toolchain.

Grobid [4] is a machine learning library for structure extraction from PDFs and other raw documents, reference extraction and parsing from PDFs, and parsing of reference strings given in plain text. Grobid uses pdfalto[8] to extract the structure of PDFs. Pdfalto is a fork of pdf2xml[9] and depends on xpdf.[10] The already mentioned Grobid dataset, used for training Grobid, also contains only a couple of footnotes which makes this library not suitable for us.

ParsCit [15] and its successor Neural ParsCit [5] take plain text as input and focus on the task of reference string parsing. ParsCit can do one more step and identify dedicated reference sections whereas Neural ParsCit is solely used for reference string parsing. ParsCit uses human-engineered features and is no longer under development. Neural ParsCit is a machine learning-based approached and is trained on scientific articles in English. SciWING [16] is a toolkit for scientific document processing. The model for reference string parsing is based on the architecture of Neural ParsCit with added ELMo embeddings [17]. These approaches are promising. But the focus on English and unsuitable training data prohibit the usage out of the box.

3. Methods

In this section, we describe the dataset we used for our analysis and our process for the development of the annotation guidelines and the annotation itself.

3.1. Dataset

We use an existing corpus of scholarly works as a starting point for our analysis. The corpus consists of 44 scholarly works interpreting *Die Judenbuche* by Annette von Droste-Hülshoff [18] and 49 scholarly works interpreting *Michael Kohlhaas* by Heinrich von Kleist [19]. The texts were originally annotated in the ArguLIT project [20] using TEI/XML [21]. The corpus contains annotations of footnotes, and annotations of all direct quotations. The quotations are assigned a type, for example, quotations from the primary literary work, other literary works, or other scholarly works. However, there are no annotations of literature references or segments of references.

[7] https://excite.informatik.uni-stuttgart.de/
[8] https://github.com/kermitt2/pdfalto
[9] https://github.com/kermitt2/pdf2xml
[10] https://www.xpdfreader.com/

3.2. Annotation

Our initial analysis of existing datasets and tools led us to the conclusion that a more in-depth analysis of the differences between literature references in literary studies and other fields is needed in order to get a better understanding of the challenges. Our annotations extend existing TEI/XML files and we therefore continue to annotate in this format. We use the Oxygen XML Editor[11] for the annotations. For our analysis, we alternated between writing/refining the annotation guidelines and annotating, where one person was responsible for writing the guidelines and the other person for the annotations (see Section 4.1).

4. Results

In this section, we present our annotation guidelines and the challenges we encountered during our analysis. The challenges are illustrated by giving examples which are taken from six scholarly works [22, 23, 24, 25, 26, 27] and are translated from German to English. Whenever possible, relevant parts are highlighted.

4.1. Annotation Guidelines and Results

Our annotation guidelines in German are available online.[12] Our annotation efforts resulted in eight fully annotated scholarly texts and two partially annotated texts. In the eight fully annotated texts, 229 complete references and 398 cross references were annotated with segments.[13] Listing 1 shows an annotated example in TEI/XML. To shorten the example, names are removed.

The example illustrates one of the challenges, namely the usage of *Id.*, and how such a case can be annotated. The example also gives an idea of the complexity of the modelling and annotation task. Often, there are multiple mentions of pages which need different types of annotations and can be hard to differentiate. Overall, we found that the difficulty of the annotation process can not be attributed to one specific issue but rather lies in the combination of a number of factors, such as the ones shown in this and the subsequent examples.

4.2. Examples for Challenges

Here, we present examples to identify and illustrate challenges and requirements for annotation and automatic extraction and segmentation. The examples are divided into two categories: *general* challenges and challenges related to *internal* references.

4.2.1. General

Example 1 shows a typical reference which gives a general idea of the complexity the references can have. The reference is build up of three levels: contained work (**red**), multi-volume work

[11] https://www.oxygenxml.com/xml_editor.html
[12] https://hu.berlin/cfe_guidelines
[13] The annotated scholarly works can currently not be shared due to copyright restrictions.

```
<bibl xml:id="bibl_Hoffmann">
  <bibl type="analytic">
    <author xml:id="author_Hoffmann">[...]</author>
    <title>Lebens-Ansichten des Katers Murr</title>.
  </bibl> In:
  <author corresp="#author_Hoffmann">Id.</author>:
  <bibl type="monogr">
    <title>Sämtliche Werke in 6 Bänden</title>. Edited by
    <editor>[...]</editor> and <editor>[...]</editor>.
    Vol. <biblScope unit="volume">5</biblScope>.
  </bibl>
  <bibl type="monogr">
    <title>Lebens-Ansichten des Katers Murr; Werke 1820-1821</title>.
    Ed. by <editor>[...]</editor>.
    <pubPlace>Frankfurt a.M.</pubPlace><date>1992</date>
    pp. <biblScope unit="page" from="9" to="458">9-458</biblScope>,
    here p. <citedRange unit="page" from="429" to="429">429</citedRange>
  </bibl> (ed. C.W.).
</bibl>
```

Listing 1: TEI/XML annotation example

(**blue**), and volume (**green**). It contains two year specifications and two page references. The internal reference *Id.* is discussed in more detail in Example 6.

(1) Ernst Theodor Amadeus Hoffmann: **Lebens-Ansichten des Katers Murr**. In: Id.: **Sämtliche Werke in 6 Volumes**. Edited by Hartmut Steinecke and Wulf Segebrecht. Vol. 5: **Lebens- Ansichten des Katers Murr; Werke 1820-1821**. Ed. by Hartmut Steinecke. Frankfurt a.M. 1992, pp. 9-458, here p. 429 (ed. C.W.).

Normally, the first quotation from the primary literary work is followed by a footnote which gives the full reference and explains the reference style in natural language (**red**). This is then followed by the specific reference for that quote (**blue**) (cf. Example 2). Extracting references that are embedded in such natural language descriptions is not supported by existing systems.

(2) The Droste texts are quoted according to Annette von Droste-Hülshoff. Historical-critical edition. Works, correspondence. Ed. by Winfried Woesler. Tübingen (Niemeyer) 1978 ff. with **HKA, volume name and page number**. Here: **HKA V, p. 3**.

Example 3 shows a reference with a report year (**red**). This can be difficult for automatic extraction as it can easily be confused with the publication date (**blue**).

(3) Cf. Jutta Linder: Strafe oder Gnade. Zur Judenbuche der Droste. Droste-Jahrbuch 3, **1991-1996**. Paderborn **1997**, pp. 83-114.

Example 4 references a letter with a description (**red**) but there is no actual title and there can be multiple letters and only the page (**green**) together with the specific edition (**blue**) indicates the exact letter.

(4) **Letter to Jenny von Laßberg**, **HKA IX**, **p. 96**.

Example 5 shows a reference that is divided between the running text (5 a) and a footnote (5 b). This is a particularly challenging example none of the existing systems can handle as they assume references to either be in a dedicated section or a footnote, but not divided between different parts of the text.

(5) (a) [...] The first time he was mentioned, he was said to have Hitler's features, for example in 1937 by Jean Cassou.[15]

(b) [15] In: Helmut Sembdner (Ed.), Heinrich von Kleist's Nachruhm, Bremen 1967, p. 455.

4.2.2. Internal references

Internal references are references between footnotes, references to other references, and references between individual segments of references. No existing system can resolve these references.

Example 6 shows a reference where the author of the article (**red**) is referenced with *id.* (**blue**) as the author of the book.

(6) **Richard Alewyn**: Origins of the detective novel, in: **id.**: Probleme und Gestalten, Essays, Frankfurt/Main 1974, pp. 341-360, here p. 350.

Example 7 shows a reference to an earlier mentioned work by giving the author (**red**) and a footnote (**blue**).

(7) **Rölleke** [**Note 1**], p. 420.

Example 8 shows a reference to the previous reference using *Ibid.* (**red**).

(8) **Ibid.**, p. 39.

Example 9 shows a reference to two earlier mentioned works by giving the author (**red**) and two footnotes (**blue**).

(9) **Wittkowski** [**Notes 5 and 6**] as a whole.

Example 10 shows a reference where the author refers to his own works by using the natural language expression *my discussion*.

(10) For more details, see **my discussion** in Kleist-Jahrbuch (1985) p. 170 ff. In the following, I refer back to my remarks in this review several times.

5. Discussion

In this paper, we have shown how literature references in literary studies differ from references in other disciplines. These differences result in a number of challenges.

First, there is basically few to no training data for machine learning approaches available. Existing datasets are tailored towards STEM or related fields and mostly target English articles. As our examples illustrate, these datasets are not suitable for our use case. Literary studies scholars make heavy use of footnotes and existing datasets contain few to no footnotes. This makes existing systems unreliable for detection of reference strings. Segmentation is also difficult as references are more complex, contain internal references, are more diverse, less structured, and more language-dependent, for example, using natural language descriptions. For the same reasons, automatic generation and manual annotation of training corpora are also more difficult. References contain more natural language expressions which makes them more language-dependent, less structured and less generalizable which in turn makes automatic generation harder. Adding to that, there are no styles available that could be used for automatic generation of references. The higher complexity makes the already complex and time consuming process of manual annotation even more difficult and therefore error prone and time consuming. To our knowledge, none of the existing systems can resolve internal references such as the ones shown in Section 4.2.2.

Current systems assume that reference detection and segmentation can be done as a two step process. Our analysis has shown that this might not be the case for literature references in literary studies. We argue that a third step of resolving internal references is necessary and that the steps need to interact. This means that a system should not assume individual references but rather look at the text as a whole and the references as a network. Joint modelling could profit from utilizing language models. This in turn would mean that a lot more training data is required. It might be possible to use existing datasets as a starting point to generating more training data by using existing databases of scholarly works and exchanging fields in templates which are either manually generated or taken from existing datasets. On top of that, in order to mimic our examples, the reference strings need to be embedded in natural language text.

Acknowledgments

Parts of this research were funded by the German Research Foundation (DFG) priority programme (SPP) 2207 *Computational Literary Studies* project *What matters? Key passages in literary works* (grant no. 424207720). We would like to thank the project's student assistant Gregor Sanzenbacher for his annotation work.

References

[1] A. Cox, What are communities of practice? A comparative review of four seminal works, Journal of Information Science 31 (2005) 527–540.

[2] S. Doerfel, R. Jäschke, G. Stumme, Publication analysis of the formal concept analysis community, in: F. Domenach, D. Ignatov, J. Poelmans (Eds.), Formal Concept Analysis,

volume 7278 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin/Heidelberg, 2012, pp. 77–95. doi:10.1007/978-3-642-29892-9_12.

[3] F. Jannidis, S. Martus, L. Konle, J. Kreutel, Was verändert sich eigentlich? Korpusanalytisch basierte Wissenschaftsgeschichte der germanistischen praxis am beispiel der deutschen vierteljahrsschrift für literaturwissenschaft und geistesgeschichte, in: Digitale Literaturwissenschaft, Stuttgart, 2019. In print.

[4] Grobid, https://github.com/kermitt2/grobid, 2008–2021. arXiv:1:dir:dab86b296e3c3216e2241968f0d63b68e8209d3c.

[5] A. Prasad, M. Kaur, M.-Y. Kan, Neural ParsCit: a deep learning-based reference string parser, International Journal on Digital Libraries 19 (2018) 323–337. doi:10.1007/s00799-018-0242-1.

[6] A. Hosseini, B. Ghavimi, Z. Boukhers, P. Mayr, EXCITE – a toolchain to extract, match and publish open literature references, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2019, pp. 432–433. doi:10.1109/JCDL.2019.00105.

[7] S. Anzaroot, A. McCallum, A new dataset for fine-grained citation field extraction, in: ICML Workshop on Peer Reviewing and Publishing Models (PEER), 2013.

[8] M. Körner, B. Ghavimi, P. Mayr, H. Hartmann, S. Staab, Evaluating reference string extraction using line-based conditional random fields: A case study with german language publications, in: M. Kirikova, K. Nørvåg, G. A. Papadopoulos, J. Gamper, R. Wrembel, J. Darmont, S. Rizzi (Eds.), New Trends in Databases and Information Systems, Springer International Publishing, Cham, 2017, pp. 137–145.

[9] G. Colavizza, M. Romanello, Annotated References in the Historiography on Venice: 19th–21st centuries, 2017. doi:10.5334/johd.9.

[10] M. Grennan, M. Schibel, A. Collins, J. Beel, Giant: The 1-billion annotated synthetic bibliographic-reference-string dataset for deep citation parsing, in: 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science, 2019, pp. 101–112.

[11] D. Thai, Z. Xu, N. Monath, B. Veytsman, A. McCallum, Using bibtex to automatically generate labeled data for citation field extraction, in: Automated Knowledge Base Construction, 2020. doi:10.24432/C5F592.

[12] Z. Boukhers, S. Ambhore, S. Staab, An end-to-end approach for extracting and segmenting high-variance references from pdf documents, in: 2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL), 2019, pp. 186–195. doi:10.1109/JCDL.2019.00035.

[13] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, Ł. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335. doi:10.1007/s10032-015-0249-8.

[14] D. Tkaczyk, P. Szostek, L. Bolikowski, GROTOAP2 - the methodology of creating a large ground truth dataset of scientific articles, D Lib Mag. 20 (2014).

[15] I. Councill, C. L. Giles, M.-Y. Kan, ParsCit: an open-source CRF reference string parsing package, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/166_paper.pdf.

[16] A. Ramesh Kashyap, M.-Y. Kan, SciWING– a software toolkit for scientific document processing, in: Proceedings of the First Workshop on Scholarly Document Processing,

Association for Computational Linguistics, Online, 2020, pp. 113–120. doi:10.18653/v1/2020.sdp-1.13.
[17] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 2227–2237. doi:10.18653/v1/N18-1202.
[18] A. v. Droste-Hülshoff, Die Judenbuche, Insel Verlag, Frankfurt am Main, 1979. URL: https://www.projekt-gutenberg.org/droste/judenbch/index.html.
[19] H. v. Kleist, Michael Kohlhaas, in: M. Holzinger (Ed.), Werke und Briefe in vier Bänden, CreateSpace Independent Publishing Platform, 1978, pp. 7–113. URL: http://www.zeno.org/nid/2000516902X.
[20] S. Winko, 2017–2020, The making of plausibility in interpretive texts. Analyses of argumentative practices in literary studies, URL: https://gepris.dfg.de/gepris/projekt/372804438?language=en.
[21] TEI Consortium, eds., 2022, TEI P5: Guidelines for electronic text encoding and interchange, version 4.4.0, URL: https://www.tei-c.org/Guidelines/P5/.

Referenced Scholarly Works

[22] K. Krauss, Das offene Geheimnis in Annette von Droste-Hülshoffs „Judenbuche“, Zeitschrift für deutsche Philologie 114/4 (1995) 542–559.
[23] V. D. Huszai, Denken Sie sich, der Mergel ist unschuldig an dem Morde. Zu Doste-Hülshoffs Novelle „Die Judenbuche“, Zeitschrift für deutsche Philologie 116/4 (1997) 481–499.
[24] D. Grathoff, „Michael Kohlhaas“, in: W. Hinderer (Ed.), Kleists Erzählungen, Stuttgart, 1998, pp. 43–66.
[25] G. Scholdt, Kleists „Michael Kohlhaas“ als Modell eines Aufruhrs, in: H. Jung (Ed.), Das Recht und die schönen Künste. Heinz Müller-Dietz zum 65. Geburtstag, Baden-Baden, 1998, pp. 115–131.
[26] C. Weder, Die (Ohn-)Macht der Objekte. Romantische Dinge zwischen Magie und Profanität in Heinrich v. Kleists „Michael Kohlhaas“ und E. T. A. Hoffmanns „Der Zusammenhang der Dinge“, in: C. Holm, G. Oesterle (Eds.), Schläft ein Lied in allen Dingen? Romantische Dingpoetik, Würzburg, 2011, pp. 145–163.
[27] G. Greve, „[...] ich habe Euch ein schweres Gewissen zu danken“. Eine psychoanalytische Interpretation der „Judenbuche“, in: G. Greve, H. E. Harsch (Eds.), Annette von Droste-Hülshoff aus psychoanalytischer Sicht, Tübingen, 2003, pp. 11–33.