Journal of Open
Humanities Data

Colavizza, G and Romanello, M 2017 Annotated References in the Historiography on Venice: 19th–21st centuries. *Journal of Open Humanities Data*, 3: 2. DOI: https://doi.org/10.5334/johd.9

DATA PAPER

Annotated References in the Historiography on Venice: 19th–21st centuries

Giovanni Colavizza and Matteo Romanello
École Polytechnique Fédérale de Lausanne, Digital Humanities Laboratory, CH
Corresponding author: Giovanni Colavizza (giovanni.colavizza@epfl.ch)

We publish a dataset containing more than 40'000 manually annotated references from a broad corpus of books and journal articles on the history of Venice. References were considered from both reference lists and footnotes, include primary and secondary sources, in full or abbreviated form. The dataset comprises references from publications from the 19th to the 21st century. References were collected from a newly digitized corpus and manually annotated in all their constituent parts. The dataset is stored on a GitHub repository, persisted in Zenodo, and it is accompanied with code to train parsers in order to extract references from other publications. Two trained Conditional Random Fields models are provided along with their evaluation, in order to act as a baseline for a parsing shared task. No comparable public dataset exists to support the task of reference parsing in the humanities. The dataset is of interest to all working on the domain of reference parsing and citation extraction in the humanities.

Keywords: Bibliographic data; Reference parsing; Citations; History; Venice
Funding Statement: The project is supported by the Swiss National Fund, with grants 205121_159961 and P1ELP2_168489.

Context

Citation indexes, such as Google Scholar, the Web of Science and Scopus, are one of the main literature retrieval tools available to modern scholars. They rest on by-now reasonably reliable large-scale reference parsers. Nevertheless, the disciplines traditionally part of the humanities are still poorly covered by citation indexes of any sort [8], something that both hinders the work of humanists and the understanding of the humanities as scholarly disciplines [1], not to mention their evaluation [4]. A key aspect of the problem is the lack of citation data, especially for local publications not in English, and for non-article publication such as scholarly monographs. The availability of citation data depends on the technical challenge of reference parsing and extraction from literature in the humanities.

Reference parsing does not exist in isolation, but depends on the digital availability of publications beforehand, and its ultimate results rest on the possibility to disambiguate any extracted reference and link it back to the identifier of the resource it points to. Open challenges to the former step are copyright, digitization and classification of publications, the main open challenge to the latter step is the absence of global repositories of metadata on the sources of humanists – especially sensible for archival materials, whose meta-ecosystems are less integrated than library catalogs.

Reference parsing poses a set of challenges in itself, which are of two kinds:

- The inherent complexity and variety of referencing practices in the humanities, both at the syntactic and semantic levels. Such variety is mostly due to disciplinary traditions, to the use of footnotes as a textual space in itself, and the variety of cited sources.
- The lack of annotated data with sufficient coverage in two critical areas: locality (of language and scholarly practice) and time (going backwards at least to the 19th century, when modern academic scholarship starts).

These two challenges make reference parsing in the humanities not intrinsically different than for the sciences, simply more involved. Several projects already exist which specifically aim at providing frameworks to extract reference data also from, or specifically from humanities' publications [5, 7, 10].

The manually annotated dataset of references released here is part of the Linked Books project[1], whose goal is to develop an in-depth approach to the problem of indexing humanities' publications via citations. The project only considers a field in historiography, the history of Venice, but does so by considering local historiography and all the modern period of the discipline (19th century to nowadays). The core idea of the project is to involve research libraries in a collaborative and distributed digitization and indexation process, by developing and providing the necessary IT infrastructure. The dataset being released was produced by librarians working for the Linked Books project during the period 2014 to 2016. Its use is to power the reference parsing and extraction modules of the platform in use for the daily operations of the project. To the best of our knowledge, no comparable dataset has been published yet.

This release is directed towards practitioners in the domain of reference parsing, with the hope that it could be of use to enrich their datasets. It is also for all interested into this specific machine learning task, with the hope that they can improve on the results here presented. Lastly, it is meant to contribute and encourage a better integration of datasets and technical tools in this domain.

Methods

The main characteristic of this dataset is its provenance, namely the corpus of publications from which it was extracted: a mixed set of monographs and journal article, with special attention to local publications in non-English languages. Secondly, it contains references to any possible source cited by historians, over a very long period of time, thus the annotation taxonomies were refined with a bottom-up approach. Thirdly, it contains references from both reference lists and footnotes, including abbreviated references which are commonplace in the humanities. Lastly, the dataset is used to train parsers using a standard technique in this domain, Conditional Random Fields.

Steps

Annotation was conducted using Brat [11][2]. It proceeded by page: all the references of a randomly picked page are annotated. If a reference spans two pages, both pages are entirely annotated. A first testing period was needed to stabilize the annotation taxonomy, whose resulting annotations have been discarded thereafter. The main challenge encountered during annotation is the presence of outlier tags: rarely occurring, yet sufficiently distinct as to warrant a category on their own. This is especially true for unpublished primary sources, whose tag variety is greater than published materials. Outlier tags need to be taken into account for automated parsing. After annotation, all annotations are consolidated and exported for further use.

Sampling strategy

The selection of the corpus of publications from which to extract references to annotated is described in detail elsewhere [3]. The rationale was to select: recent monographs and the complete archive of specific journals, at the aid of library catalog, scholarly bibliographies and domain experts. The result was a first collection of 1922 monographs and 3 journals: *Ateneo Veneto*, *Archivio Veneto* and *Studi Veneziani*, for a total of 552 issues. After digitization and OCR, the latter done using ABBYY FineReader Corporate v12, a second sampling was conducted for annotation, namely:

- 196 monographs were randomly picked and their reference lists completely annotated.
- 144 journal issues were randomly picked and a set of references were annotated from their footnotes (a minimum of two contiguous pages for each article in the issue, leaving annotators to select pages dense in references). The first issue was published in the year 1866, the last in 2013, in order to cover all periods of interest and variations in referencing practices therein.

Quality Control

The quality of the annotations is guaranteed by the joint work of annotators, who were working at the same time in the same room, thus consulting each other on problematic choices. No double-keyed annotation on a subset of the data has been conducted at this date.

Annotation taxonomies

The main annotation distinction was made between generic and specific tags, or whole references and their components. *Generic tags* included the distinction between primary sources (such as archival documents), secondary sources (books) and meta sources (secondary sources published within a container source, such as journal articles or contributions in edited volumes). This classification choice is motivated by a) the difference in their components (specific tags) and b) the needs of the look-up module in our pipeline (which matches a reference with a unique identifier in an internal or external repository, such as a library catalog, in order to define a citation. Different external resources are used for any given generic category). *Specific tags* include instead all the possible components of the three classes of references mentioned above, such as author, title and publication year for secondary sources, archive, archival reference and archival unit for primary sources. More examples are given in **Tables 1** and **2**. The full taxonomy is available in the GitHub repository associated with this article.

Dataset description

The annotated dataset is given as a zipped JSON file within a repository containing extra details and code to train parsing models.

Object name

LinkedBooksReferenceParsing v1.1.

Format names and versions

JSON, Python 3.

Creation dates

2014 to 2016.

Dataset Creators

Giovanni Colavizza, Matteo Romanello, Martina Babetto and Silvia Ferronato.

Language

English. Contents are in a variety of languages, mainly Italian, English, French, German, Spanish and Latin.

License

CC BY Attribution 4.0 International.

Repository name

GitHub and Zenodo.

Publication date

13/05/2017.

Statistics and contents

Basic statistics of the dataset.

Statistic	Value
Total annotations	198'839
Generic annotations	41'071
Specific annotations	157'768
Generic from monographs (reference lists)	11'360
Generic from journal articles (footnotes)	29'711
Annotated documents over the whole database	14%
Avg. annotated pages per annotated document	17

Parsers trained with the dataset

Two parsers were trained using the annotated dataset. First, a parser assigns specific tags on the full-text of new publications (model 1: *citation parsing*), secondly, another parser assigns generic and begin-end tags to the same full-text, relying on the results of the first parser (model 2: *citation extraction and classification*). Both parsers use Conditional Random Fields (CRF), a standard technique for text parsing tasks [6][3]. The interested reader can find an introduction to CRFs in [12]. Preliminary parsing results on subsets of the dataset are already reported elsewhere, including a more detailed description of the challenges encountered, features used and ablation tests conducted in order to select the best performing combination of features [2, 3]. The full code is provided for replication in the repository associated to this article.

Table 1: Evaluation of model 1 on the test set, with best parameters.

Key	Tag	Precision	Recall	F1-score	Support
0	<empty>	0.778	0.856	0.815	23'144
1	Tomo	0.660	0.445	0.532	440
2	Archive_lib	0.910	0.744	0.818	2'095
3	Foliation	0.927	0.878	0.902	706
4	Numbered_ref	0.594	0.590	0.592	1'799
5	Box	0.587	0.728	0.650	942
6	Publisher	0.876	0.698	0.777	2'397
7	Date	0.804	0.773	0.788	2'790
8	Series	0.715	0.772	0.742	2'377
9	Folder	0.535	0.244	0.335	726
10	Volume	0.699	0.448	0.546	2'007
11	Author	0.893	0.865	0.879	13'543
12	Column	0.500	0.486	0.493	70
13	Cartulation	0.955	0.891	0.922	1'376
14	Publicationnumber-year	0.793	0.747	0.770	1'631
15	Title	0.886	0.922	0.903	48'566
16	Conjunction	0.498	0.446	0.470	688
17	Publicationspecifications	0.439	0.346	0.387	1'433
18	Archivalreference	0.769	0.760	0.764	5'248
19	Abbreviation	0.842	0.719	0.776	1'389
20	Ref	0.454	0.368	0.406	307
21	Registry	0.822	0.606	0.698	883
22	Pagination	0.909	0.921	0.915	7'232
23	Year	0.912	0.873	0.892	3'822
24	Attachment	0.598	0.353	0.444	1'484
25	Publicationplace	0.856	0.845	0.850	3'555
26	Filza	0.954	0.940	0.947	284
	Avg / total	**0.837**	**0.840**	**0.836**	**128'794**

Both models were trained as follows. First, the annotated dataset was consolidated in order to group similar and under represented tags under the same tag. Details are given in the repository's README file. Afterwards, for each model 10% of the relevant annotations were kept aside for validation, of the remaining 90%, 25% is considered as test and 75% as train data. Using a quasi-Newton gradient descent method (L-BFGS), there are two main parameters in CRFs: c1 for L1 and c2 for L2 regularizations, respectively. The provided models use the following parameters:

- Model 1, c1: 0.07; c2: 0.378.
- Model 2, c1: 0.09; c2: 0.447.

Another relevant choice for the CRF models is the dependency window to consider, which was set to two tokens before and after the one under consideration.

The evaluation of both models with best parameters, on the test set is given in **Tables 1** and **2**, to be read along with the confusion matrices in **Figure 1**.

Both models perform acceptably well if one considers the most important tasks they have. For model 1, these regard being correct on the most discriminative (and represented) tags such as author, title or archival reference. For model two, this entails getting the extraction task correctly (begin-end), something more important than getting the classification correctly, which is performed

Table 2: Evaluation of model 2 on the test set, with best parameters.

Key	Tag	Precision	Recall	F1-score	Support
0	b-meta (begin)	0.730	0.819	0.772	1745
1	b-primary	0.837	0.821	0.829	2067
2	b-secondary	0.826	0.828	0.827	5694
3	e-meta (end)	0.729	0.801	0.763	1754
4	e-primary	0.823	0.818	0.820	2074
5	e-secondary	0.838	0.832	0.835	5770
6	i-meta (in)	0.823	0.870	0.846	34091
7	i-primary	0.884	0.900	0.862	14203
8	i-secondary	0.881	0.844	0.862	48960
9	o (out)	0.957	0.943	0.950	30730
	Avg / total	**0.875**	**0.874**	**0.874**	**147088**

Table 3: 5-fold validation and final validation scores.

Model	5-fold average f1 score	Validation precision	Validation recall	Validation f1 score
1: parsing	0.834	0.829	0.832	0.829
2: extraction and classification	0.930	0.908	0.908	0.908

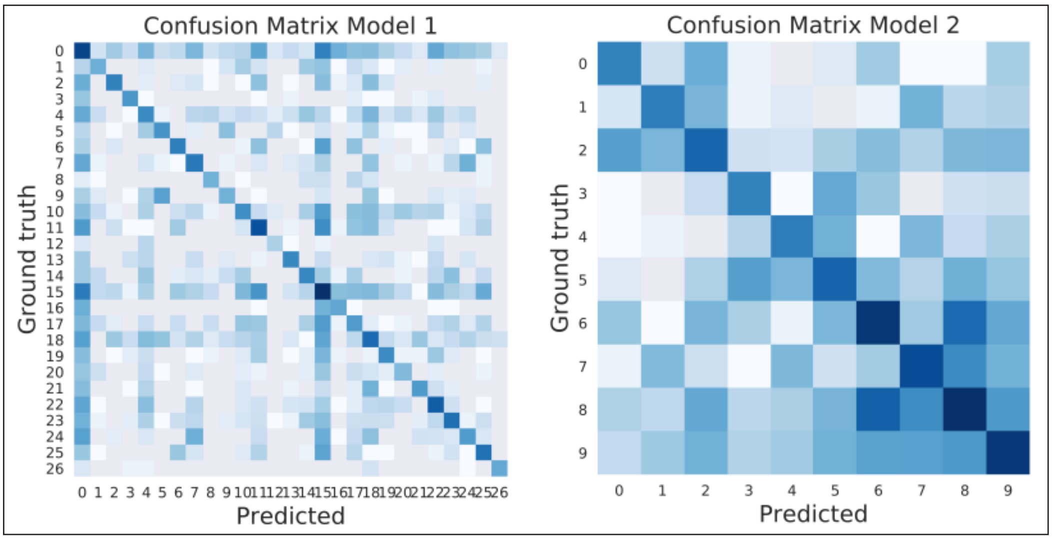

Figure 1: Confusion matrices for model 1 and 2. The numbers of rows and columns correspond to the keys in the above tables. The color map goes from grey (zero) to dark blue (max of examples).

decently given that most errors entail miss-classifications at the classification but not the begin-end task.

The 5-fold validation and final validation results are given in **Table 3**, by considering models now trained on all but validation data.

Reuse potential and future work

The main use of this dataset is to train new or enrich existing reference extraction tools with more data, of a kind normally difficult and costly to find. The dataset might be of use also to teachers and interested researchers willing to experiment machine learning techniques in order to improve upon our results: the code is shared in order to encourage not only replication but especially improvement.

The dataset comes with a number of limitations, most notably its domain specificity. As it happens it is unknown, and an interesting open question, to what extent this annotated dataset can perform well on similar tasks but for different contents. To the extent possible, the provided models are largely language-independent, due to the fact that the corpus already contains a variety of languages.

We plan to focus next on the release of larger quantities of both manually and automatically produced annotations as linked data. We suggest that two immediate open challenges for the community are: sharing and federating annotated data for reference parsing in the humanities under unique standards; subsequently developing general parsers which could be reliably applied to a variety of different collections.

Repository location

GitHub: https://github.com/dhlab-epfl/LinkedBooksReferenceParsing
Zenodo: http://doi.org/10.5281/zenodo.579679

Notes

[1] http://dhlab.epfl.ch/page-127959-en.html.

[2] http://brat.nlplab.org/.

[3] The implementations used were CRFsuite [9] and sklearn-crfsuite https://github.com/TeamHG-Memex/sklearn-crfsuite.

Acknowledgements

This work would not have been possible without the help of several Italian institutions of culture. In alphabetical order we thank: the *Ca' Foscari* University Library System and the *Ca' Foscari* University Humanities Library (BAUM), the Central Institute for the Union Catalogue of Italian Libraries and Bibliographic Information (ICCU), the European Library of Information and Culture (BEIC), the *Istituto Veneto di Scienze Lettere ed Arti*, the *Marciana* Library, the State Archive of Venice.

Competing Interests

The authors have no competing interests to declare.

References

1. **Ardanuy, J** 2013 Sixty years of citation analysis studies in the humanities (1951–2010). *Journal of the American Society for Information Science and Technology, 64*: 1751–1755. DOI: https://doi.org/10.1002/asi.22835
2. **Colavizza, G** and **Kaplan, F** 2015 On Mining Citations to Primary and Secondary Sources in Historiography. In *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-It 2015 3–4 December 2015, Trento*, pp. 94–99. DOI: https://doi.org/10.4000/books.aaccademia.1439
3. **Colavizza, G, Romanello, M** and **Kaplan, F** 2017 The References of References: A Method to Enrich Humanities Library Catalogs with Citation Data. *International Journal on Digital Libraries*: 1–11. DOI: https://doi.org/10.1007/s00799-017-0210-1
4. **Hammarfelt, B** 2016 Beyond Coverage: Toward a Bibliometrics for the Humanities. In Ochsner, M, Hug, S E and Daniel, H-D (Eds.), *Research Assessment in the Humanities*. Cham: Springer International Puublishing, pp. 115–131. DOI: https://doi.org/10.1007/978-3-319-29016-4_10
5. **Kim, Y M, Bellot, P, Faath, E** and **Dacos, M** 2011 Automatic annotation of bibliographical references in digital humanities books, articles and blogs. In *Proceedings of the 4th ACM workshop on Online books, complementary social media and crowdsourcing, ACM*, pp. 41–48. DOI: https://doi.org/10.1145/2064058.2064068
6. **Lafferty, J, McCallum, A** and **Pereira, F** 2001 Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of ICML*, pp. 282–289.
7. **Lopez, P** 2009 GROBID: combining automatic bibliographic data recognition and term extraction for scholarship publications. In: *Research and Advanced Technology for Digital Libraries*, Springer, pp. 473–474. DOI: https://doi.org/10.1007/978-3-642-04346-8_62
8. **Mongeon, P** and **Paul-Hus, A** 2016 The Journal Coverage of Web of Science and Scopus: A Comparative Analysis. *Scientometrics, 106*(1): 213–28. DOI: https://doi.org/10.1007/s11192-015-1765-5
9. **Okazaki, N** 2007 CRFsuite: a fast implementation of Conditional Random Fields (CRFs). Available at www.chokkan.org/software/crfsuite.
10. **Romanello, M** 2015 *From Index Locorum to Citation Network: an Approach to the Automatic Extraction of Canonical References and its Applications to the Study of Classical Texts* (PhD dissertation). King's College London. http://hdl.handle.net/11858/00-1780-0000-002A-4537-A.
11. **Stenetorp, P, Pyysalo, S, Topi, G, Ohta, T, Ananiadou, S** and **Tsujii, J** 2012 BRAT: a web-based Tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, EACL '12*, pp. 102–107.
12. **Sutton, C** and **McCallum, A** 2011 An Introduction to Conditional Random Fields. *Foundations and Trends in Machine Learning, 4*(4): 267–373. DOI: https://doi.org/10.1561/2200000013

How to cite this article: Colavizza, G and Romanello, M 2017 Annotated References in the Historiography on Venice: 19th–21st centuries. *Journal of Open Humanities Data*, 3: 2. DOI: https://doi.org/10.5334/johd.9

Published: 21 November 2017

]u[*Journal of Open Humanities Data* is a peer-reviewed open access journal published by Ubiquity Press

OPEN ACCESS