SMAuC - The Scientific Multi-Authorship Corpus

Philipp Sauer
Leipzig University

Janek Bevendorff
Bauhaus-Universität Weimar

Lukas Gienapp
Leipzig University

Wolfgang Kircheis
Leipzig University

Erik Körner
Leipzig University

Benno Stein
Bauhaus-Universität Weimar

Martin Potthast
Leipzig University

Abstract

With an ever-growing number of new publications each day, scientific writing poses an interesting domain for authorship analysis of both single-author and multi-author documents. Unfortunately, most existing corpora lack either material from the science domain or the required metadata. Hence, we present SMAuC, a new metadata-rich corpus designed specifically for authorship analysis in scientific writing. With more than three million publications from various scientific disciplines, SMAuC is the largest openly available corpus for authorship analysis to date. It combines a wide and diverse range of scientific texts from the humanities and natural sciences with rich and curated metadata, including unique and carefully disambiguated author IDs. We hope SMAuC will contribute significantly to advancing the field of authorship analysis in the science domain.

1 Introduction

Authorship analysis at its heart deals with discriminating between different writing styles or attributing writings styles to specific authors. With early approaches dating back to the 19th century (Koppel et al., 2009), the issue has been addressed with a wide range of ever more sophisticated methods founded in linguistics, psychology, and, increasingly important, computer science.

Yet despite decades of research in computational authorship analysis, scientific papers remain a challenge as they are comparatively short and often form at least a collection if not a homogeneous blend of the different writing styles contributed by their co-authors. Additionally, many scientific disciplines require researchers to adapt to a specific set of stylistic requirements and leave little room for personal expression. Extracting personal stylistic qualities from such multi-author documents is therefore particularly challenging.

A first step in tackling this problem is a thorough and rigorous stylometric comparison of both monographs and multi-author documents involving the same author. If stylistic characteristics of an author from a number of monographs can be found in documents with multiple authors, it becomes easier to identify passages which were most likely written by that author or to which they contributed. In an effort to aid researchers in developing and improving computational authorship analysis methods on scientific texts in particular, we release SMAuC, a collection of 3,356,686 scientific papers of both multi-author and single-author origin that are enriched with faceted and disambiguated metadata. To our knowledge, it is the largest collection of text compiled explicitly for the purpose of authorship analysis to date, not only in the science domain, but also for general stylistic or stylometric inquiry.

In the following, we review existing corpora of openly available scientific texts (Section 2) and describe the dataset curation process (Section 3), followed by a qualitative and quantitative discussion of the dataset contents, the collected metadata, and the corpus as a whole (Section 4). Finally, we summarize key data and give an outlook of potential applications for the corpus (Section 5).

2 Related Work

While research on authorship has yielded several datasets in the scientific field, very few are available or can be reproduced: Payer et al. (2015) collect 6,872 conference papers in an effort to develop methods for deanonymization of scientific publications. Sarwar et al. (2018) aggregate 2,573 papers from the arXiv preprint service to conduct multi-author attribution; for the same purpose, Rexha et al. (2016) collect 6,144 articles from the PubMed database. Boumber et al. (2018) introduce and publicly release the MLPA-400 dataset, which consists of 400 scientific publications. Larger collections are available for general-purpose authorship analysis, for example the PAN-20 Style Change Detec-

Conditions applied	**Number of documents**	
CORE	123,988,821	*(100.00%)*
↪ full texts	9,835,064	*(7.93%)*
↪ text language filtering	6,531,442	*(5.27%)*
↪ OAG matching	3,508,509	*(2.82%)*
↪ text quality assurance	3,356,686	*(2.70%)*

Table 1: Dataset curation process with number of documents remaining after each step. Percentages are relative their original counts.

tion Corpus (Zangerle et al., 2020) consisting of approximately 23,000 stack exchange postings.

Contrary to this, a very limited number of corpora encompassing large amounts of scientific texts is indeed available, yet none of these were specifically designed for authorship analysis: Soares et al. (2018) use a self-constructed corpus of roughly 30,000 scientific documents in Portuguese, English, and Spanish for research on automated translation. Citron and Ginsparg (2015) present a corpus of 757,000 scientific texts for text reuse detection extracted from arXiv.org. Gipp et al. (2014) introduce a dataset of 234,591 articles from approximately 975,000 authors extracted from the PubMed Central Open Access Subset, a large collection of biomedical full texts, many of which are available with an open access license. A corpus of 1.14M paper full texts was used by both Beltagy et al. (2019) and Ammar et al. (2018). The papers were obtained from Semantic Scholar and originate from computer science and biomedical research.

While readily available, all of the corpora mentioned above exhibit one or several shortcomings: They are either very small and therefore only of limited use to large-scale authorship attribution, they lack the metadata required for authorship analyses, or they are too narrow in scope, i.e., limited to one scientific domain only. This necessitates the creation of a new large-scale dataset specifically curated for this purpose, which covers a larger variety of scientific disciplines with detailed metadata.

3 Dataset Creation

SMAuC is created by merging data from two sources: the CORE database (Knoth et al., 2011; Knoth and Zdrahal, 2012), a large collection of metadata and full texts of open access scientific publications, and the Microsoft Open Academic Graph (OAG, Sinha et al., 2015), a large, openly accessible heterogeneous knowledge graph based on scientific articles, authors, and institutions.

As a basis for our dataset, we used the CORE database dump from 2018-03-01[1]. It comprises 123M metadata items, of which 85.6M items have abstracts and 9.8M items have the full texts. Each item represents a single scientific paper or book. The OAG serves as an additional source for identifying and disambiguating the authors and fields of study of the publications. We rely on Version 2 of the OAG (Hu et al., 2020)[2] with 179 million nodes and 2 billion edges.

Table 1 illustrates the four-step selection process we applied to all entries in the final corpus: (1) From the CORE corpus, we selected all entries with full texts and (2) filtered these for English-language articles. (3) We matched the selected subset with their corresponding OAG metadata to obtain unique author and fields of study information. (4) Finally, we applied certain text quality heuristics for ensuring a high-quality extraction.

From the 123M CORE entries, we extracted a total of 9.8M entries with available full texts. Although CORE specifies a language flag, it is only present in some entries. We added missing language flags using a standard fastText (Joulin et al., 2016, 2017) language detection model. For that, the texts were split into five parts of equal length of which at least four needed to be English. 6,531,442 entries remained after this step.

In the third step, entries were merged with metadata in the OAG. An official mapping between the two already exists (Version 2019-04-01[1]), yet it contains only 655K of the 6.5M English entries. Furthermore, the DOIs (as given in CORE and OAG) were not accurate in some cases. If we simply used these DOIs as keys, we could be seeing false positive and false negative matching errors. To reduce the number of matching errors, we defined two extra matching criteria of which at least one had to be met to count as match: (1) The DOIs of both entries had to be identical and both titles had to have a Levenshtein distance of less than 10% the length of the shorter title. (2) The titles and years of publication had to be identical and at least one author name had to appear in both entries with a low Levenshtein distance as detailed above.

With this method, we were able to match OAG metadata for 3.5M CORE entries, a significant improvement over the official CORE-OAG-mapping. Yet, postprocessing the metadata was required in

[1] `https://core.ac.uk/services/dataset`

[2] `https://www.microsoft.com/en-us/research/project/open-academic-graph/`

(a)

Document Type	Count
Single author w/o multi author	711,471
Single author w/ multi author	261,629
Multi author w/o single author	1,481,106
Multi author w/ single author	894,945
No author information	7,535
Total	3,356,686

(b)

Length	Total	Single author		Multi author	
≤ 3,000	39,300	13,680	(1.41%)	25,567	(1.07%)
– 5,000	96,067	32,059	(3.29%)	63,832	(2.69%)
– 50,000	2,273,246	467,844	(48.07%)	1,799,435	(75.73%)
– 250,000	771,756	301,975	(31.03%)	468,473	(19.72%)
> 250,000	176,317	157,542	(16.19%)	18,744	(0.79%)
Total	3,356,686	973,100	(100.00%)	2,376,051	(100.00%)

Table 2: **(a)** Counts for all types of documents and their total; **(b)** Number of documents in the corpus by text length in characters and document type with percentage per row. Documents with no author information are omitted. Length values refer to the raw texts including tables, captions, and appendices.

some cases. The fields of study given in the OAG per publication are of varying granularity (e.g., 'humanities' as a whole vs. 'chemical solid-state research' as a subfield of chemistry). To establish a standardized, hierarchical scheme, we manually mapped the annotated disciplines to the *DFG Classification of Scientific Disciplines and Research Areas* (DFG, 2016). The mapping was carried out manually by three persons at very high agreement. Cases of disagreement were discussed internally and subsequently unified. The final three-level hierarchy includes *disciplines* (Engineering Sciences, Humanities & Social Sciences, Life Sciences, Natural Sciences), *research areas* (e.g. Chemistry, Medicine, Mechanical and Industrial Engineering, . . .), and *fields* (e.g. Educational Research, Condensed Matter Physics, Zoology, . . .).

In the final quality assurance step, the full texts of all entries were cleaned by removing markup and all non-ASCII characters, converting texts to lower case letters and collapsing runs of whitespace characters. Then, two heuristics were used to eliminate texts of sub-par quality: (1) Cleaned texts with a length below 2,000 characters (approximately one printed page) were excluded. (2) Cleaned texts were split at sentence boundaries into three equally sized chunks and the fastText language detection model was once again applied to each part individually. If fastText considered a part to be English with more than 60% confidence, this part was accepted as English. An entry was excluded if more than one of the three parts was not classified as English. This repeated round of language classification was to further ensure that only English texts remain, since the first (coarse) round was performed on the uncleaned texts. The small number of entries (152,000) removed in this step suggests that the coarse filtering already excluded most non-English text reliably.

4 Corpus Description

This section describes the structure, format, and key properties of SMAuC. The corpus is distributed in the form of multiple line-delimited JSON files, each containing 100,000 entries. The entries all have identifiers (DOI, CORE, OAG) and detailed meta information about the publication (title, abstract, citation count, reference), its authors (name and OAG identifier), the general discipline and field of study, as well as the full text from CORE.

Corpus Size and Composition. Table 2a details the composition of the corpus itemized by document type. Publications can be split into two fundamental categories: (1) monographs (i.e., single-author) and (2) multi-author (i.e., collaborative) publications. By investigating author relations, we can further differentiate each of the two into subtypes, for a total of four document types: (1) single-author publications whose authors have not participated in any multi-author publications; (2) single-author publications whose authors appear in at least one multi-author document; (3) multi-author publications whose authors have not written any monographs; and (4) multi-author publications with at least one author who has written at least one additional monograph. In addition, for a very small subset of documents, no author information is available. These texts will not be immediately useful for attributing the texts to specific authors, though they may still be useful material if larger collections of text from specific research areas are needed, e.g., for general comparison between sub-corpora of the humanities and sciences.

Overall, the corpus contains fewer monographs than multi-author documents. Of these, only a minority of monograph authors have participated in multi-author publications and vice versa. Documents with no author information are rare. The

Research Area	SA	MA	A	TL
Engineering Sciences	55,015	375,206	3	28,467
Humanities	58,317	199,926	3	37,224
Life Sciences	48,723	715,218	5	32,616
Natural Sciences	147,024	651,076	3	26,103

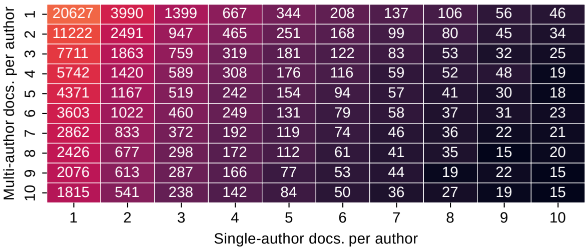
Table 3: Single author documents (**SA**), multi author documents (**MA**), median authors per document (**A**) and median text length (**TL**), by research area.

total document count exceeds previous datasets on authorship analysis by a big margin.

Text Lengths. The corpus comprises a wide range of texts of different lengths: from very short articles of just a few pages to long book-sized entries. Table 2b lists document counts for different length bins. The bins were chosen as approximate character counts for (1) abstract papers (less than one page), (2) short papers (1–2 pages), (3) essay-length papers (up to 10 pages), (4) long papers (up to 50 pages), (5) and books or dissertations (more than 50 pages). Most papers in the corpus are between 5,000 and 50,000 characters in length (2 and 20 pages). Multi-author publications are shorter on average than single-author publications. A sizable portion of publications exceed lengths of 250,000 characters, most of which are monographs. These seem to be mainly individual dissertations and less often collaborative book publications. Manual spot checks confirmed this assumption.

Academic Disciplines. Fields of study annotations are available for approximately 1.7M entries in the corpus. Table 3 lists document counts and key statistics per discipline, reflecting different publishing practices across fields. For example, the median text length is higher in the humanities compared to the natural sciences, while the median author count is highest for the life sciences. The relative proportion of single- and multi-author documents also differs per discipline, yet in all disciplines, sufficient amounts of either type are present to conduct authorship analyses.

Author Information. Establishing reliable author relations between documents is paramount for use as a ground-truth for authorship analysis. To this end, the OAG provides disambiguated and unique author IDs. Of particular interest to us are authors involved in both single- and multi-author publications. In total, 5,664,224 unique authors are present in the corpus. Of these, 670,566 appear exclusively in single-author publications and 4,868,263 exclusively in multi-author documents. The remaining 125,395 authors, who appear in both types of documents, are thus of particular interest. Figure 1 shows author counts over the number of single-author and multi-author publications. Unsurprisingly, the vast majority of authors appear in at most one or two single- or multi-author publications, respectively. More than five documents per author in either category are increasingly rare.



	1	2	3	4	5	6	7	8	9	10
1	20627	3990	1399	667	344	208	137	106	56	46
2	11222	2491	947	465	251	168	99	80	45	34
3	7711	1863	759	319	181	122	83	53	32	25
4	5742	1420	589	308	176	116	59	52	48	19
5	4371	1167	519	242	154	94	57	41	30	18
6	3603	1022	460	249	131	79	58	37	31	23
7	2862	833	372	192	119	74	46	36	22	21
8	2426	677	298	172	112	61	41	35	15	20
9	2076	613	287	166	77	53	44	19	22	15
10	1815	541	238	142	84	50	36	27	19	15

Figure 1: Total author count over the number of single-author and multi-author publications per author. Publication counts beyond 10 are omitted (35,178 authors).

5 Conclusion

We introduce SMAuC, the largest available corpus for authorship analysis in the scientific domain. It encompasses over 3.3M documents and detailed, standardized metadata including author and field-of-study annotations. The corpus allows to select subsets of texts according to numerous criteria, each in itself still met by a significant number of documents. Selecting only very short texts is just as possible as picking entire volumes; including only authors with a high number of monographs will still generate subsets with several thousands of texts. Even selecting only multi-author texts for which individual writing style analyses are supported by additional monographs, leaves a subset of more than 70,000 documents. If smaller subsets are sufficient, it is also possible to combine constraints, e.g. select all multi-author texts only from the humanities with additional monographs for all authors. The corpus allows for compiling a myriad of interesting subsets tailored to a wide range of very specific research questions in authorship analysis, particularly in, but not restricted to, the science domain. SMAuC is available on Zenodo.[3]

[3] http://doi.org/10.5281/zenodo.7289788

Ethics Statement

Our dataset compiles contemporary writing from the domain of science ("papers") with the purpose of studying the capabilities of authorship analysis technology in dealing with scientific papers and the challenges that arise from multi-author documents.

Ethical considerations for datasets in general relate to four main areas of concern (Peng et al., 2021), three of which are relevant to this paper: (1) privacy of the individuals included in the data, (2) effects of biases on downstream use, and (3) dataset usage for dubious purposes. We therefore took into account a consensus on best-practices for ethical dataset creation (Mieskes, 2017; Leidner and Plachouras, 2017; Gebru et al., 2021).

Ad (1). An anonymization or pseudonymization of the papers in our corpus is virtually impossible, since they are publicly available, and a querying the original CORE/OAG data would reveal the author(s) of every enclosed paper. By partaking in the scientific discourse, however, any published paper becomes part of science's legacy, which is open to everyone to make it their subject of analysis, scrutiny, and mining. This is especially true for articles under an open-access license, where consent to the creation of derivative works, public archiving, and mining is implied.

Ad (2). Stylometry is particularly prone to confounding variables such as text domain, genre, or audience (Koolen and van Cranenburgh, 2017; Bevendorff et al., 2019; Bischoff et al., 2020), which replicates to downstream tasks. No explicit measures for preventing such biases in the data can be taken given the wide variety of authorship-related tasks that can be studied. Rather, we opt to include as much data and metadata as possible to enable researchers to derive their own datasets for their specific tasks, allowing them to address confounding factors individually. The dataset strives for transparency and extendability by documenting its creation process and by retaining references to the original data sources.

Ad (3). We deem the overall abuse potential of the corpus low, particularly in comparison to what is already possible today with the OAG. Yet, as a further precaution, access to the data will be granted on a per-request basis via Zenodo for academic use only.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the Literature Graph in Semantic Scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, Volume 3*, pages 84–91, New Orleans - Louisiana. ACL.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. ACL.

Janek Bevendorff, Matthias Hagen, Benno Stein, and Martin Potthast. 2019. Bias Analysis and Mitigation in the Evaluation of Authorship Verification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6301–6306, Florence, Italy. Association for Computational Linguistics.

Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, Ben Thies, Matthias Hagen, Efstathios Stamatatos, Benno Stein, and Martin Potthast. 2020. The Importance of Suppressing Domain Style in Authorship Analysis. *CoRR*, abs/2005.14714.

Dainis Boumber, Yifan Zhang, and Arjun Mukherjee. 2018. Experiments with convolutional neural networks for multi-label authorship attribution. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. ELRA.

Daniel T. Citron and Paul Ginsparg. 2015. Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences of the United States of America*, 112(1):25–30.

Deutsche Forschungsgemeinschaft DFG. 2016. DFG classification of scientific disciplines, research areas, review boards and subject areas. Accessed 2021-05-27.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM*, 64(12):86–92.

Bela Gipp, Norman Meuschke, and Corinna Breitinger. 2014. Citation-based plagiarism detection: Practicability on a large-scale scientific corpus. *Journal of the Association for Information Science and Technology*, 65(8):1527–1540.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2704–2710. ACM.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *ArXiv preprint*, abs/1612.03651.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. ACL.

Petr Knoth, Vojtech Robotka, and Zdenek Zdrahal. 2011. Connecting repositories in the open access domain using text mining and semantic data. In *Research and Advanced Technology for Digital Libraries*, volume 6966, pages 483–487.

Petr Knoth and Zdenek Zdrahal. 2012. CORE: Three Access Levels to Underpin Open Access. *D-Lib Magazine*, 18(11/12).

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by design: Ethics best practices for natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. Association for Computational Linguistics.

Margot Mieskes. 2017. A quantitative study of data in the NLP community. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 23–29, Valencia, Spain. Association for Computational Linguistics.

Mathias Payer, Ling Huang, Neil Zhenqiang Gong, Kevin Borgolte, and Mario Frank. 2015. What you submit is who you are: A multimodal approach for deanonymizing scientific publications. *IEEE Trans. Inf. Forensics Secur.*, 10(1):200–212.

Kenneth Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual.*

Andi Rexha, Stefan Klampfl, Mark Kröll, and Roman Kern. 2016. Towards a more fine grained analysis of scientific authorship: Predicting the number of authors using stylometric features. In *Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval co-located with the 38th European Conference on Information Retrieval (ECIR 2016), Padova, Italy, March 20, 2016*, volume 1567 of *CEUR Workshop Proceedings*, pages 26–31.

Raheem Sarwar, Chenyun Yu, Sarana Nutanong, Norawit Urailertprasert, Nattapol Vannaboot, and Thanawin Rakthanmanon. 2018. A scalable framework for stylometric analysis of multi-author documents. In *Database Systems for Advanced Applications - 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part I*, volume 10827 of *Lecture Notes in Computer Science*, pages 813–829. Springer.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246, New York, NY, USA. ACM.

Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A Large Parallel Corpus of Full-Text Scientific Articles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. ELRA.

Eva Zangerle, Maximilian Mayerl, Günther Specht, Benno Stein, and Martin Potthast. 2020. Overview of the Style Change Detection Task at PAN 2020. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.