SMAuC – The Scientific Multi-Authorship Corpus

Janek Bevendorff
Bauhaus-Universität Weimar

Philipp Sauer
Leipzig University

Lukas Gienapp
Leipzig University

Wolfgang Kircheis
Leipzig University

Erik Körner
Leipzig University

Benno Stein
Bauhaus-Universität Weimar

Martin Potthast
Leipzig University

ABSTRACT

The rapidly growing volume of scientific publications offers an interesting challenge for research on methods for analyzing the authorship of documents with one or more authors. However, most existing datasets lack scientific documents or the necessary metadata for constructing new experiments and test cases. We introduce SMAuC, a comprehensive, metadata-rich corpus tailored to scientific authorship analysis. Comprising over 3 million publications across various disciplines from over 5 million authors, SMAuC is the largest openly accessible corpus for this purpose. It encompasses scientific texts from humanities and natural sciences, accompanied by extensive, curated metadata, including unambiguous author IDs. SMAuC aims to significantly advance the domain of authorship analysis in scientific texts.

KEYWORDS

Authorship Analysis, Multi-Authorship

1 INTRODUCTION

Authorship analysis focuses on distinguishing writing styles or attributing them to specific authors. Originating in the 19th century [16], this field has developed through various methods from linguistics, psychology, and notably, computer science. However, computational authorship analysis still struggles with scientific papers, as they tend to be relatively short and often contain a mix of co-authors' writing styles. Furthermore, many disciplines impose strict stylistic guidelines, limiting personal expression. Consequently, extracting individual stylistic traits from multi-authored documents remains a significant challenge.

To address the problem of multi-author authorship analysis, a comprehensive stylometric comparison of monographs and multi-author documents involving the same author is required. Identifying an author's stylistic features in multi-author documents allows for more accurate identification of their contributions. To facilitate corresponding advancements in computational authorship analysis for scientific texts, we introduce SMAuC, a dataset of 3,356,686 scientific papers with both single and multi-author origins, accompanied by detailed and disambiguated metadata. To our knowledge, this is the largest dataset compiled specifically for authorship analysis, suitable for research on scientific texts and broader stylometric investigations.[1,2]

We first examine current publicly accessible scientific text corpora (Section 2), then detail our dataset creation methodology (Section 3). Subsequently, we provide a qualitative and quantitative analysis of the dataset, its metadata, and the entire corpus (Section 4). Lastly, we discuss corpus applications (Section 5), followed by and ethical self-assessment.

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. [20] collect 6,872 conference papers in an effort to develop methods for de-anonymization of scientific publications. Sarwar et al. [23] aggregate 2,573 papers from the arXiv preprint service to conduct multi-author attribution; similarly, Rexha et al. [22] collect 6,144 articles from the PubMed database for the same purpose. Boumber et al. [5] introduce and publicly release the MLPA-400 dataset, which consists of 400 scientific publications. Larger multi-author collections are available for other domains, for example the PAN-20 Style Change Detection Corpus [26] consisting of approximately 23,000 stack exchange postings, combining questions with answers to form multi-author documents.

Of the already very limited number of corpora that include large amounts of scientific texts, none were specifically designed for authorship analysis: Soares et al. [25] use a self-constructed corpus of roughly 30,000 scientific documents in Portuguese, English, and Spanish for research on automated translation. Citron and Ginsparg [6] present a corpus of 757,000 scientific texts for text reuse detection extracted from arXiv. Gipp et al. [9] introduce a dataset of 234,591 articles 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 1.14 million paper full texts was used by both Ammar et al. [1] and Beltagy et al. [2]. The papers were obtained from Semantic Scholar and originate from computer science and biomedical research.

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 [13, 14], a large collection of metadata and full texts of open access scientific publications, and the Microsoft Open Academic Graph (OAG, [24]), a large, heterogeneous knowledge graph based on scientific articles, authors, and institutions.

[1]Code: https://github.com/webis-de/JCDL-23
[2]Download: http://doi.org/10.5281/zenodo.7289788

Table 1: Dataset curation process with number of documents remaining after each filtering step. Percentages relative to full CORE.

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

As a basis for our dataset, we used the CORE[3] database dump from 2018-03-01. 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 [10] with 179 million nodes and 2 billion edges.[4]

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 [11, 12] language detection model. The texts were split into five parts of equal length of which at least four needed to be English. After applying this filter, 6,531,442 entries remain.

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. Using these DOIs as keys could result in 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 a match: (1) The DOIs of both entries had to be identical and both titles had to have a Levenshtein distance of less than 10% of 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. However, postprocessing the metadata was required in 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* [7]. 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 Sci., Humanities & Social Sci., Life Sci., Natural Sci.), *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 white-space 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 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. Each entry has identifiers (DOI, CORE, OAG) and detailed meta information about the publication (title, abstract, citation count, reference), its authors (name and OAG identifier), the discipline and field of study, and 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) single-author (i.e., monographs) and (2) multi-author (i.e., collaborative) publications. By investigating author relations, we can further differentiate each of the two into sub-types, 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 publication; (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 comparisons between sub-corpora of the humanities and sciences.

The corpus contains fewer monographs than multi-author documents. Only a minority of monograph authors have participated in multi-author publications and vice versa. Documents with no author information are rare. The total document count exceeds previous datasets on authorship analysis.

[3]https://core.ac.uk/services/dataset
[4]https://www.microsoft.com/en-us/research/project/open-academic-graph/

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.

(a)

Document Type	Count
Single author without multi author	711,471
Single author with multi author	261,629
Multi author without single author	1,481,106
Multi author with 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 3: Single-author and multi-author documents, median authors per document (Med.) and median text length, by research area.

Research Area	Single author	Multi author	Med.	Length
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

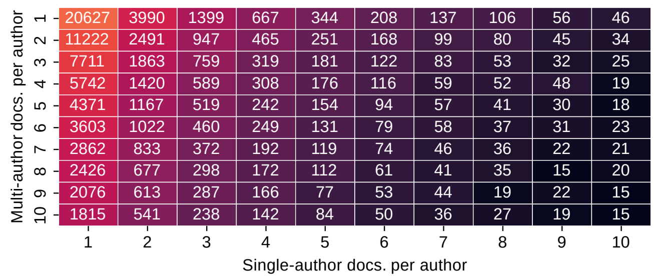

Figure 1: Author count over number of single-author and multi-author publications per author. Counts beyond 10 are omitted (35,178 authors).

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 (one or two pages), (3) essay-length papers (up to 10 pages), (4) long papers (up to 50 pages), and (5) 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.

The 973,100 monographs in the corpus can be attributed to 795,000 different authors. The majority of them is represented with just one monograph (92.02%). This also means that 8% of the authors have written more than 24% of the monographs in the corpus. For those 8% it will be possible to extract individual stylistic features based on a larger set of texts. For a total of 125,395 authors, both single-author and multi-author documents are present, amounting to a total of 1.15M documents. Most of these authors have written a small number of single-author and an even smaller number of multi-author documents.

Corpus File Description. The corpus is distributed as `xz`-compressed file in `JSONL` format, where records are each encoded as a single `JSON` string per line. The corpus can thus be efficiently stream-processed without requiring to inflate the corpus file. An example record with associated keys, datatypes, description is specified in Table 4.

Table 4: Schema for a single data record with field names, datatypes, descriptions, and randomly drawn example.

	Key	Datatype	Description	Example
IDs	`core_id`	`int`	ID of the publication in the CORE corpus	2461603
	`doi`	`string`	DOI of the publication	10.1103/PhysRevD.63.123512
	`download_url`	`string`	URL of fulltext PDF	http://arxiv.org/abs/hep-ph/0012097
	`mag_ids`	`array of int`	Microsoft Academic Graph ID	[2064413872]
Metadata	`authors`	`array of (int, string)`	Author list with ID and name per author	[(2238602696, Arttu Rajantie), ...]
	`doc_type`	`string`	Type of the publication	Journal
	`fields_of_study`	`array of string`	Array of field names	[Particles, Nuclei and Fields, ...]
	`publisher`	`string`	Publisher of the publication	The American Physical Society
	`venue`	`(int, string) tuple`	Venue ID and name	(173952182, Physics Letters B)
	`issue`	`string`	Issue the publication appeared in	12
	`volume`	`string`	Volume the publication appeared in	63
	`year`	`int`	Year of publication	2001
	`n_citation`	`int`	Number of citations	29
Text	`title`	`string`	Title of the publication	Electroweak preheating on a lattice
	`abstract`	`string`	Abstract text	In many inflationary models, large ...
	`fulltext`	`string`	Parsed fulltext	In many inflationary models, large ...

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 relevant subsets tailored to very specific research questions in authorship analysis, particularly in, but not restricted to, scientific texts.

The core element of the corpus are the 1,144,915 documents for which monographies and multiauthor documents from the same authors are available, which could be of particular interest for authorship analysis in the context of multiauthorship. Future research with the corpus could include the application of well-tested authorship analysis methods like Unmasking [17] on the domain of academic texts. Those methods can now also be tested on their ability to detect writing styles of authors extracted from their monographies in documents co-authored by other researchers.

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 [21], 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 [8, 18, 19].

Ad (1). An anonymization or pseudonymization of the papers in our corpus is virtually impossible, since they are publicly available, and querying for 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 [3, 4, 15], 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. It is a collection of texts and metadata obtained from publicly available sources, assembled with respect to their terms and conditions.

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

[1] 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 *Proc. of NAACL'18*. ACL, 84–91.

[2] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proc. of EMNLP-IJCNLP'19*. ACL, 3615–3620.

[3] Janek Bevendorff, Matthias Hagen, Benno Stein, and Martin Potthast. 2019. Bias Analysis and Mitigation in the Evaluation of Authorship Verification. In *Proc. of the 57th Annual Meeting of the ACL*. ACL, 6301–6306.

[4] 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 (2020).

[5] Dainis Boumber, Yifan Zhang, and Arjun Mukherjee. 2018. Experiments with Convolutional Neural Networks for Multi-Label Authorship Attribution. In *Proc. of LREC'18*. ELRA.

[6] Daniel T. Citron and Paul Ginsparg. 2015. Patterns of text reuse in a scientific corpus. *PNAS* 112, 1 (2015), 25–30.

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

[8] 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 (2021), 86–92.

[9] Bela Gipp, Norman Meuschke, and Corinna Breitinger. 2014. Citation-based plagiarism detection: Practicability on a large-scale scientific corpus. *JASIST* 65, 8 (2014), 1527–1540.

[10] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *Proc. of WWW'20*. ACM, 2704–2710.

[11] 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 (2016).

[12] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proc. of EACL'17*. ACL, 427–431.

[13] 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*, Vol. 6966. 483–487.

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

[15] Corina Koolen and Andreas van Cranenburgh. 2017. These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proc. of the First ACL Workshop on Ethics in Natural Language Processing*. ACL, 12–22.

[16] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *JASIST* 60, 1 (2009), 9–26.

[17] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research* 8 (2007), 1261–1276. https://doi.org/10.5555/1314498.1314541

[18] Jochen L. Leidner and Vassilis Plachouras. 2017. Ethical by Design: Ethics Best Practices for Natural Language Processing. In *Proc. of the First ACL Workshop on Ethics in Natural Language Processing*. ACL, 30–40.

[19] Margot Mieskes. 2017. A Quantitative Study of Data in the NLP community. In *Proc. of the First ACL Workshop on Ethics in Natural Language Processing*. ACL, 23–29.

[20] 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 (2015), 200–212.

[21] Kenneth Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating dataset harms requires stewardship: Lessons from 1000 papers. In *Proc. of NeurIPS'21*.

[22] 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 *Proc. of ECIR'16*, Vol. 1567. 26–31.

[23] 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 *Proc. of DASFAA'18*, Vol. 10827. Springer, 813–829.

[24] 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 *Proc. of WWW'15*. ACM, 243–246.

[25] Felipe Soares, Viviane Moreira, and Karin Becker. 2018. A Large Parallel Corpus of Full-Text Scientific Articles. In *Proc. of LREC'18*. ELRA.

[26] Eva Zangerle, Maximilian Mayerl, Günther Specht, Benno Stein, and Martin Potthast. 2020. Overview of the Style Change Detection Task at PAN 2020. In *Proc. of CLEF'20*, Vol. 2696. 11 pages.