University of Leipzig
Faculty of Mathematics and Computer Science
B.Sc. Digital Humanities

Enabling Authorship Analysis in Scientific Texts

Bachelor's Thesis

Philipp Sotirios Sauer
Born Jul 23, 1994 in Miltenberg

Matriculation Number 3060255

1. Referee: Jun. Prof. Dr. Martin Potthast

Submission date: August 16, 2021

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Leipzig, August 16, 2021

. . . .
Philipp Sotirios Sauer

Abstract

The task of authorship analysis includes the identification of people according to their individual writing style, author profiling and the dissection of multi-author documents by contributions of different individuals. This Bachelor's thesis makes a contribution to the enabling of computational methods in authorship analysis for scientific texts by constructing the **S**cientific **A**uthorship **A**nalysis **D**ocuments (SAAD), a new corpus of scientific documents.

Existing corpora in the field of authorship analysis are diverse, but few of them include scientific texts. The corpus presented here is supposed to fill this gap and enable future research on the suitability of existing authorship analysis or verification methods for the realm of scientific research and multiauthor documents. The corpus is designed to include a broad range of texts from different scientific disciplines, ranging from stylistically more diverse texts from humanities to more standardized and stylistically homogenous texts from natural sciences.

Texts and necessary metadata are extracted and merged from two existing datasets and combined in an enriched and standardized new dataset. The dataset is focused on providing monographies and multiauthor-documents from the same authors, to allow the analysis of individual writing styles and the way those become visible in different contexts of multi-authorship. The corpus will be made accessible in a graph database, together with a tool to specify and extract subsets of the data. The tool is supposed to provide an easy and intuitive way of exploring the corpus and will enable its users to find texts tailored for their own specific research-questions.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Thesis Motivation and Problem Description

Authorship analysis describes the task of discriminating between the writing styles of authors via a broad range of methods based in linguistics, psychology and, increasingly important, computer science. Its methods have evolved since the 19th century, [Koppel et al., 2009] and have become increasingly more sophisticated ever since. Modern methods like Unmasking [Koppel and Schler, 2004] or Impostors Method [Seidman, 2013] are no longer limited to appliance on large amounts of text only, but can provide insights on the authorship of rather short texts with high levels of confidence.
However, up to today only little research has been published on the suitability of those methods for scientific texts. Scientific publications and papers pose several challenges for authorship analysis: they are often short, and the presence of direct or indirect citation between authors may obfuscate their individual styles. Additionally, many scientific disciplines require researchers to adapt to a very specific set of stylistic requirements and leave little room for personal expression. Especially for multiauthor documents on their own, extracting personal stylistic qualities is a nearly impossible challenge. The general assumption that the first two or maybe first three authors listed will have written the biggest part of the paper may hold in many empirical cases, but on an individual case-basis, it remains an assumption that cannot be verified directly. Comparing monographies and multiauthor documents from the same author may be a first step to tackle this problem. The writing style extracted from a number of monographies can then be traced in documents with multiple authors, so it might become possible to identify passages which are more or less likely to be written by this specific author. However, to test this, a sufficiently large collection of monographies and multiauthor documents by the same authors must be available.

As already mentioned, practices of scientific writing may vary immensely between scientific fields. While some philosophers are well known for their distinct style, which may additionally be of great importance for the expression of their thoughts, only few physicists will be remembered for the way they wrote their work. This may affect the outcome of authorship verification methods, and corpora for this task should be able to make differences between scientific fields visible.

After all, the lack of research on scientific authorship analysis may have less to do with the challenges it poses, and more with the availability of corpora of scientific texts to test and adapt existing methods on. Few large corpora of scientific texts exist, and those existing may not be ideal for specific research purposes. The increasing availability and popularity of open access research over the last years opens up an opportunity to fill this gap. This thesis in particular will explore the potential of the CORE dataset (see chapter 3) for the construction of a large corpus of scientific texts with rich metadata and make the attempt to build such a corpus, to enable future research on authorship analysis for scientific texts. The resulting corpus will be presented under the name **S**cientific **A**uthorship **A**nalysis **D**ocuments (SAAD).

1.2 Requirements Towards the Corpus

The corpus to be constructed should fulfil some requirements based on the specific challenges of authorship analysis on scientific texts. Most importantly, it should collect a combination of monographies and multiauthor documents for a large number of authors. The authors must be identifiable and possible misattributions of authors with identical names, changing names and different ways of spelling must be minimized or completely avoided at best. Entries must include machine-readable full texts of the articles (not only abstracts), and must be written in English. English articles unsurprisingly constitute the largest part of the CORE dataset, and the construction of a balanced multilingual corpus would require more detailed knowledge on the other languages to be included, so non-english texts will be left out for the work presented here.

However, a multilingual corpus may be beneficial in other research scenarios, and CORE may be a suitable basis for this as especially many Spanish and Russian articles are included in the dataset.

Furthermore, the papers should include information on the research field they belong to. This information will be mapped and standardized according to the classification of scientific research by the German Research Association to ensure a classification by a common schema.

1.3 Working Process and Thesis Structure

The corpus presented in this work is based on two existing, large scale datasets. Its starting point is the CORE-dataset, the second large source is the Microsoft Academic Graph. CORE provides metadata for 123 Million open access research papers, and full texts for 9.8 Million of those. The information extracted from CORE is enriched by information from the Microsoft Academic Graph and PDF-extracted full texts from the Webis-group. The generated dataset is to be stored in a graph-based Neo4J database. A web-interface for interacting with the database is provided, which allows to search and visualize entries from the corpus in a browser and extract subsets of the data via a search-and-export function.

Chaper 2 of this thesis will give a brief overview on current research about authorship verification and existing corpora, to contextualise the research questions the corpus is designed to answer. Chapter 3 will describe the datasources, their strengths and problems and how they can be combined to make authorship analysis tasks on the data possible. The process of combining those will be discussed in Chapter 4. The resulting corpus and its contents are the topic of Chapter 5, followed by a brief introduction into the details and features provided by the browser based corpus-tool in Chapter 6.

Chapter 2

Theoretical Background

There is a variety of research topics which texts from the presented corpus might be used for. Those may be tasks as basic as a statistical comparison of word frequencies among different scientific fields, or more complex tasks like testing automated classifiers for scientific topics of texts. However, the two main research areas the corpus was developed for are authorship analysis and plagiarism detection. Challenges in those fields that should be adressed with the help of the corpus will be discussed in the following chapter.

2.1 Authorship Analysis

2.1.1 Basics

Authorship analysis can be used as an umbrella term for the tasks for authorship verification, authorship attribution/identification, and authorship profiling. [Brocardo et al., 2013] Authorship profiling in this context means inferring sociodemographic characteristics of an unknown author from their written text, authorship identification the selection of the most likely author of a text of unknown source from a list of possible authors based on texts already attributed to them, and "[a]uthorship verification is the problem of inferring whether two texts were written by the same author." [Bevendorff et al., 2019a, p. 654] The three problems are differing in their interest and perspective but remain closely linked to each other.

The terminology may vary across different papers (e.g. Juola [2007] uses "authorship attribution" for what is described as authorship profiling by Brocardo et al.), but in this thesis, the terms will be used as defined in the above section.

First scientific approaches to authorship analysis date back as far as the 19th century, when the idea of using statistical methods to identify a stylometrical fingerprint of an author was discussed for the first time. An early measure adopted for this purpose was the relationship between word length and the relative frequency of occurrence of words of the respective length. [Koppel et al., 2009] A landmark study was published by Mosteller and Wallace [1963] using a Bayes classifier to classify the contested authorship of the Federalist papers.

2.1.2 Methods

Current methods of authorship analysis are based on statistical measures of text and machine learning techniques. There is a vast range of methods used [Stamatatos, 2009].
The unmasking approach measures how fast the ability of a variety of possible similarity measures to accurately discriminate between same- and different-author cases drops when the most characteristic features of a text are removed consecutively. Subsequently, a meta-classifier can be trained on the degradation curves to discriminate between same- and different-author curves, achieving high accuracies for book-size texts. [Koppel and Schler, 2004] The authors describe their basic idea as testing "the rate of degradation of the accuracy of learned models as the best features are iteratively dropped from the learning process." [Koppel and Schler, 2004] Though Koppel/Schler expect their method to be likely robust against changes of language or genre, as no features specific to either have to be used, they also write that this is a hypothesis that has to be confirmed still. Other studies using their method indeed demonstrate performance-reduction of unmasking with certain gernes, especially theatrical texts, that are likely to require adapted settings of the unmasking algorithm. [Kestemont et al., 2012] Similar phenomena might occur with scientific texts.
Another approach for authorship verification is the so-called Impostors Method (IM), whose "main idea is to use repeated feature subsampling methods to determine if one document of the pair allows us to select the other from among a background set of 'impostors' in a sufficiently robust manner." [Koppel and Winter, 2014, p. 178] For this, IM uses a broad range of distance measures between vector representations of texts. [Kocher and Savoy, 2017] The approach has also been adapted for comparing a single document with a set of documents [Seidman, 2013] and has been made more robust for working with short texts. [Potha and Stamatatos, 2020]
Any authorship analysis task becomes easier to solve when the size of available text is increased. This is quite intuitive, as statistic principles behind all methods will always be more reliable for large amounts of data. For shorter texts like tweets, messages or many scientific papers, adaptions are necessary. Bevendorff et al. [2019a] have introduced a generalized version of the unmasking approach by Koppel and Schler [2004] that reduces the number of required

text significantly, so authorship verification becomes possible for text fragments or very short texts. Their approach exploits the “bag of words nature of unmasking features” Bevendorff et al. [2019a, p. 255] and generates additional chunks of text from the existing material. It achieves state of the art performance and is customizable to prioritize result precision for real world forensic cases where a high confidence in the result of the method is required.
As exemplified, different methods for authorship verification, and for authorship analysis in general, may perform differently on certain types of texts, depending on their length and genre. For a reasonable judgement of their performance, they have to be tested on different benchmark corpora. Bevendorff et al. [2019b] have demonstrated possible sources of bias in corpora and how they may influence the performance of authorship verification methods if applied on small corpora. Bias in the PAN15 dataset for authorship verification allowed fairly simple methods to achieve competitive results, but completely failed on other datasets. They state that "[t]his is frequently the case when machine learning meets small data. Inadvertent properties of the data act as confounders that a learning algorithm will gladly fit onto if they are not controlled." This emphasizes the need for large corpora and diverse datasets for authorship analysis research. Bevendorff et al. [2019b, p. 6305]

2.2 Plagiarism Detection and Text Reuse

Next to authorship analysis, another research field that is strongly dependent on the type of data this thesis deals with is plagiarism detection. There are in general two approaches to automated plagiarism detection: intrinsic and extrinsic. The first one is based on identifying suspicious writing style differences, the second on comparing documents to a set of source documents [Barrón-Cedeño et al., 2010] Intrinsic plagiarism detection is closely linked to authorship analysis tasks, [Stein et al., 2011, p. 63] as the identification of "style breaches" [Foltýnek et al., 2020] is central to finding out which passages of a text may not be written by its official author.
Intrinsic approaches are limited by similar factors as authorship analysis tasks, so the length of available text is of essence. For producing reliable results, a threshold of around 35,000 words per analysed text is desirable. [Gipp et al., 2014, Stein et al., 2011] Though plagiarism also occurs in other genres of text, it is especially relevant in academic contexts, and corpora of scientific texts are needed for most tasks.

2.3 Existing, Large-scale Corpora of Scientific Texts

There is not a large number of corpora that could be of relevance here, as scientific texts haven't been the focus of a lot of research in recent years. The following section will describe existing corpora of scientific texts. Purpose and usage of these corpora vary greatly.
Yasunaga et al. [2019] present an annotated corpus for automated scientific article summarization. It contains around 1000 papers from ACL Anthology Papers from Computational Linguistics which have been manually annotated for processing. Soares et al. [2019] use a self-constructed corpus of roughly 30.000 scientific documents in Portuguese, English and Spanish, in which all documents are available in at least two of those languages. The corpus is used for research on automated translation. Corpora of larger scale that could be adapted for the research field of this thesis more easily are for example found in Citron and Ginsparg [2015]. They use a corpus of scientific texts for text reuse detection which they extracted from arXiv.org. Main fields of study present in this corpus are physics, mathematics and computer science. Their corpus consists of around 757,000 articles and covers a time-span from 1991 to 2012. Gipp et al. [2014] use a dataset with "234,591 articles by approximately, 975,000 authors from 1,972 peer-reviewed journals (as of April, 2011)." [Gipp et al., 2014, p. 1529] Their research is also focused on the extrinsic detection of plagiarism and text reuse, and the articles they use are 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. The dataset can be expected to have substantially grown since the time the data for their study was extracted, but keeps a thematically very narrow focus of biological and medicinal research.
The largest corpus described here was used by two studies. [Ammar et al., 2018, Beltagy et al., 2019] Beltagy et al. use the corpus constructed by Ammar et al. one year earlier for training a language model for scientific texts. They use 1.14M papers originating from Semantic Scholar, around one fifth from computer science, the remaining part from biomedical research. All the articles used contain the full text of the papers. Their dataset is a subset of a graph constructed by Ammar et al.. Their approach and the structure of the dataset is quite similar to the process that will be described later in this thesis. Ammar et al. obtained PDFs of scientific articles via “publishers (e.g., Springer, Nature), catalogs (e.g., DBLP, MEDLINE), pre-publishing services (e.g., arXiv, bioRxive)" [Ammar et al., 2018, p. 85] and web-crawling some non-specified sources. The obtained PDFs were split in a set of attributes which were automatically extracted from the PDFs, as ‘title’,’authors’,’abstract’ or

'full text'. The dataset was then processed in the form of a literature graph, containing nodes for articles, authors or publicating venues. Their final graph contains 280M nodes of different kinds, 37M of those representing single papers. From the paper, it does not become completely clear for how many of those articles full texts are available. Like the subset used by Beltagy et al., the graph contains texts from computer science and biomedical research.

For the construction of the corpus by Ammar et al., similar problems as with the corpus presented here occur. The issue of author disambiguation is discussed, and a binary classifier is used to decide whether to merge two authors with similar names. For non-abbreviated names, a higher similarity is required, while for abbreviated names a match of initials is considered sufficient.

Chapter 3

Data Sources

3.1 CORE

The CORE-Dataset is the basis for the corpus presented in this thesis. CORE is a non-profit service delivered by The Open University and the UK educational service provider Jisc. It is a large collection of metadata and full texts on open access scientific articles. Currently, metadata for a total of 207,255,818 articles can be found in the dataset. This data is collected from repositories, open access or hybrid journals. The data is constantly harvested from a total of more than 10,000 data providers, collected and made available via an API and in the form of a dataset.
By 2015, an estimated 47% of scientific publications was some form of OA [Piwowar et al., 2018], with numbers steeply increasing over the last decade. CORE tries to build on this development. Data about the papers is harvested from Open Access Repositories using the Open Archives Initiative Protocol for Metadata Harvesting [Horwood et al., 2004] which returns paper data in XML form according to Dublin Core specifications (more on Dublin Core metadata schemata in chapter 6). Subsequently, data is harmonized and made accessible through a variety of services like an API or a plugin for library services. [Knoth and Pontika, 2016, Knoth and Zdrahal, 2012, Knoth et al., 2011] The first prototype of CORE has been developed by Petr Knoth in 2010, and the dataset has been expanded since then. For this thesis, the dataset 2018-03-01 has been used. It includes 123M metadata items, 85.6M items with abstracts and 9.8M items with full text, where each item is supposed to represent a single scientific paper or book. Only full text including items have been considered for the corpus, as papers with abstracts only will not be suitable for authorship verification purposes. Beyond abstracts and full texts, CORE provides diverse metadata on the articles. The complete JSON structure of CORE entries can be found in Figure A.7.

Most importantly, a list of authors is included which is one of the basic requirements for the task to be achieved in this thesis. Also, some identifiers as DOIs and OAIs are included, as well as information on the publishing journals, XML data and fields for topics and subjects. This may at first look like all requirements for the corpus formulated in chapter 1 can already be met with the data from CORE. A list of authors is given, a language tag is present and fields of study can be derived from the 'topics' and 'subjects' fields. However, large parts of the metadata is not available for all, or, in many cases, for rather few of the entries in the dataset. Only coreId, title, authors, abstract, fulltext and year can be expected to be present in all or at least most of the entries. For example, DOIs are only available for 2.8M of the 9.8M entries with fulltexts, topics are listed for at least 6.2M and language information can be found for 2.2M. Even if all of these data-points are available, problems derive from the way they are structured.

1. Authors
   Looking at the data in more detail, it becomes visible that some common errors occur throughout the entries. Authors may be listed double, in some cases making documents with just one author appear as multiauthor documents. In other instances, the author-lists are nested, containing some of the authors as a sublist or repeating the whole list of authors as a nested list as last element. Using functions to calculate the length of the authorlist to get information about the number of authors will create misleading results in these cases, flattening all lists may result in many duplicate authors.
   Secondly, there is no ideal way here to identify and connect authors. Entries can be linked by the name of their authors, but whether those names refer to one and the same person indeed cannot be said with certainty. A similar problem may appear if an author is one time listed with his full middle name and in one case with his middle name abbreviated – in such cases, it can only be guessed if the same person wrote the paper without manual lookup and research, which cannot be accomplished for all of the 9.8M entries.

2. Language
   Language tags in CORE are helpful for selecting English texts, however they still leave out a large portion of the dataset which might contain perfectly helpful articles for the corpus. Additionally, the existing language tags are in some occasions wrong. Checking with automated language recognition tools quickly reveals articles tagged as English, despite the fact that they clearly aren't. (see chapter 4) This does not seem like a

widespread problem across the dataset, but still should be considered as a possible source of errors in the corpus.

3. DOIs
   DOIs are given for 2,846,221 of the CORE entries. 2,474,668 of those are unique, which already gives a hint towards possible inconsistencies in the data. While some of those non-unique DOIs are indeed the result of duplicate entries, many are simply incomplete and present only the beginning of DOIs, which cannot be resolved to actual documents. A sample-testing of 4,146 DOIs via the crossref.org API returned for 2,193 of those either a 404 error or a different title than the one referenced in the CORE entry.

4. Full texts
   The fulltexts in CORE are generated via automated extraction from PDF-files mostly. This difficult task can be expected to generate problematic results in some occasions. Common problems are for example texts where every character is separated by whitespace separately, or large sections of unrecognized characters. This means, for the final corpus, fulltexts must be checked to leave out as many unreadable entries as possible.

5. Topics
   Topics and subjects are given for around two thirds of the full text including entries in CORE. Unfortunately, the tags are rather unspecific at times like "Research Article" or "Original Article", or highly specific at others like "Condensed Matter - Strongly Correlated Electrons". This makes it difficult to connect the articles with a certain field of study, even in instances where they are present. The entries contain a total of 4,591,608 different topic tags. To map all of those to a field of study system manually would be impossible, some heuristic or automated analysis would be required.

3.2 Microsofts Open Academic Graph

Dealing with those problems of CORE requires additonal data. The Microsoft Academic Graph (MAG) is a large, heterogeneous graph knowledge base that contains scientific publications, authors, institutions, papers, venues, citation relations and information about conferences. [Sinha et al., 05182015] It is used for many of Microsoft's services like the Cortana personal assistant and is updated regularly. The version used for the construction of this corpus is taken

from the OAG v2 (Open Academix Graph). It contains 179 million nodes and 2 billion edges. [Hu et al., 2020] As authors are represented in the form of nodes in the MAG, the problem of identifying them is solved by the structure of the data. All documents can be linked by the edges emerging from the author-node. Even problems like name-changing due to e.g. marriages become less of an issue, as the MAG information is created using individual ORCID iDs.
The MAG-schema is quite detailed and can be found at Eide and Huang [28.05.2021]. For this thesis, only nodes of the types ‘Paper’, ‘Author’, ‘Journals’ and ‘FieldsOfStudy’ have been used. Finding the articles with full text from CORE in the MAG and linking them by the MAG-authorship-edges will be the basis for the corpus. Additionally, for around half of the entries, it allows an easier attribution of papers to their respective fields of studies and does not require deriving those from the several thousand different topic tags given in CORE.

3.3 CORE-MAG Mapping

CORE offers a mapping of its entries to the MAG dated 2019-04-01. The mapping contains corresponding MAG entries for 8.9M of the CORE entries. Considering the high number of 123M items in CORE, this is a rather low percentage and reflects in the fact that only 724,497 of the entries with full text can be found in this mapping at all, not cosidering their language or other selection features for the corpus. To match and use as many CORE entries as possible for the final corpus, a separate mapping was created. This contains around 0.5M matches from the official mapping, but could find a much larger intersection between the datasets beyond those (see chapter 4).

Chapter 4

Corpus Construction

4.1 Selection Process

Table 4.1: Selecting documents from the CORE dataset: Number of documents remaining after applying conditions

	Stages of document selection	
	Conditions applied	Number of documents
1	CORE entries with fullText tag	9,835,064
	- Non-english texts	- 3,303,622
2	English entries	6,531,442
	- No match in Microsoft Academic Graph	- 3,022,933
3	With match in MAG	3,508,509
	- Low quality full texts	- 151,823
4	Texts included in final corpus	3,356,686
	singleauthor	973,100
	multi-author	2,376,051
	no-author	7,535
	- Author has only monographs in corpus, or all authors have only multiauthor-documents in corpus	- 2,200,112
4	At least one author has both a monograph and multiauthor-document in corpus	1,156,574

Table 4.1 illustrates the selection process that lead to the entries presented in the final corpus. The following section will go through the single steps and will discuss methods used for each step. Source code can be looked up in the GitLab repository of this thesis (see section A.1).

4.1.1 Selecting Articles with Full Text

The tarball.gz File for the CORE Dataset 2018-03-01 contains each data item as a single JSON-line. All 123M entries have been extracted and only those have been unzipped that contained a 'fullText'-key in their JSON-body. This left a total of 9.8M entries.

4.1.2 Selecting all English Full Texts

In general, CORE entries have language tags that allow to select entries with a certain language. As mentioned earlier, those tags have some imperfections, and from the 9.8M entries with fulltexts, a large portion of 7.6M entries has no language tag at all. To select all English articles, the language tags were considered trustworthy in general (incorrectly tagged entries will be sorted out at a later point). Additionally, Facebook AI lab's fastText library was used for classifying the language of those texts without a specific tag. fastText offers a pretrained language recognition model for 176 different languages. It works on the basis of a linear classifier and is optimized for speed and memory usage, while staying competitive with deep learning techniques on many common tasks in terms of accuracy. [Joulin et al., 2016a,b, see]
For language recognition, the texts were split in 5 parts of equal length, for which the most likely language was chosen individually by the fastText model. Texts with two or more sections identified as non-english were eliminated from the set. The confidence the language model had in its language prediction was not yet considered in this step but will be later in the process. Splitting the text in five parts showed to be slightly beneficial over a smaller number of parts, probably due to the relatively common sections of unrecognizable characters throughout the fulltexts of CORE, which would make the whole section they belong to unrecognizable to the model. A test set of 11,367 entries with CORE language tag, among those 7,759 non-english and 3,608 english, returned the following results for the librarys fastText and langid Table 4.2.

Due to the high number of texts that had to be processed, the langid library had to be considered too slow as its accuracy benefits appeared to be insignificant. In this case, the CORE-tags, even though provenly incorrect in some cases, were used as ground-truth. A manual checkup of the false negative and positive cases showed that for many of those, the language detection actually revealed an incorrect CORE language tag. For this reason, the accuracy of the models can in fact hardly be judged based on the presented numbers, as the ground truth will be flawed in some instances. However, for the purpose of developing a heuristic for pre-selecting texts for the corpus, they are at least reassuring that in most cases, the fastText selection will find the desired texts.

Table 4.2: Comparison of python language recognition packages on 11,367 pre-tagged CORE entries

Langid vs. fastText: Accuracy and Speed			
Documents (total)	Tagged English	Tagged non-English	
11,367	3,608	7,759	
	Texts with English language tag classified non-English	Texts without English language tag classified as English	Runtime
langid	3,9%	0,3%	25m24s
fastText	4,1%	0,6%	44s

After this, around one third of the CORE entries were eliminated, leaving 6.5M entries with a high probability of being actual English texts.

4.1.3 Matching with Microsoft Academic Graph

The remaining entries had to be checked for matching entries in the Microsoft Academic Graph. The mapping provided by CORE itself contained only a small fraction of around 655K of the 6.5M English entries, so a separate matching process was necessary. Performing individual title-comparisons over a string distance measure like the Levenshtein distance for all English CORE entries with all 208M MAG entries would have been too time-consuming. As an alternative, all DOIs and titles from the CORE entries were put in a sorted set for quicker search in these sets. For any MAG entry, it was initially checked if either the DOI or the title produced a hit in those sets. Comparing the MAG entries with the DOIs and titles from the much smaller CORE set saved time as it limited the size of the sets. If so, all CORE entries with the respective title or DOI were looked up in detail and other properties were comparisond. To be considered as a match, one of the two following cases had to be fulfilled:

1. The DOIs of both entries had to be identical, and the titles of both entries had to have a sufficiently low Levenshtein distance. Matching by DOIs only produced some incorrect matchings where one (or both) of the entries had an incorrect DOI, so comparing titles as well made matches more reliable.

2. The titles of both entries had to be identical, the year of publication had to be identical and at least one author had to appear in both entries with a sufficiently low Levenshtein distance.

This reduced the overall number of necessary comparisons from at least 1.36×10^{15} to $4,18 \times 10^8$ set-lookups and around 10×10^6 individual comparisons. Matching was performed via a Jupyter notebook which can be found under notebooks/mag_matching.ipynb. After a manual control of around 300 of the generated matches, the process was repeated with slightly modified parameters, after one example had been found where an incorrect match had occurred. This had resulted from an overlap in titles, where "Meditative Reflections on Nils Christie's 'Words on Words' Through an African Lens" had been matched with "More words on words", both entries incorrectly containing the DOI of the original article "Words on words". The Levenshtein comparison after the DOI match in this case considered "More words on words" as a sufficiently similar substring of "**M**editative Reflections **o**n Nils Ch**r**isti**e**'s '**Words on Words**' Through an African Lens". After matching with an adapted threshold for substrings, a second manual lookup didn't find any incorrect matches. In the end, matching entries could be found for 3.5M of the CORE entries. This is a significant improvement over the official CORE-MAG-mapping.

4.1.4 Sort Out of Texts with Low Quality

So far, the quality of the fulltexts has only been considered as far as they were at least partially recognizable for the fastText language recognition. The final corpus however is supposed to contain only those fulltexts that allow some meaningful authorship analysis, which excludes really short texts and texts which are not recognizable as humanly written. For sorting out entries, the fulltexts were cleaned first. First, HTML or XML tags were removed, subsequently all non-ASCII characters were removed as well and all letters were converted to lowercase. Finally, all types of single or multiple whitespace were converted into a single space.
Two heuristics for sort out were applied subsequently:

1. Cleaned texts with a length of below 2000 characters (approximately 1 page printed) were excluded.

2. Texts were split up in sentences and divided in three parts containing an equally large number of sentences. Then again a pretrained fastText model was used to determine the language of those sections individually. If fastText considered a textpart to be English with more than 60% confidence, this part was accepted as English. If more than one of the three parts was not classified as English, the entry was excluded.

The number of parts to split the text into and the threshold were results of manual experiments with a small number of texts and could have been chosen differently without a large impact on the number of texts excluded. Filtering results like this left out an additional 152K documents, leaving a total of 3.4M documents.

4.1.5 Author Intersection

With the goal of exploring scientific authorship in mind, the part of the corpus of authors with both monographies and multiauthor-documents is especially relevant. After matching with the Microsoft Academic Graph, authors can be identified via their MAG-Id. For 1.1M documents, at least one author has both monographies and multiauthor-documents available in the given corpus.

4.2 Merging Sources

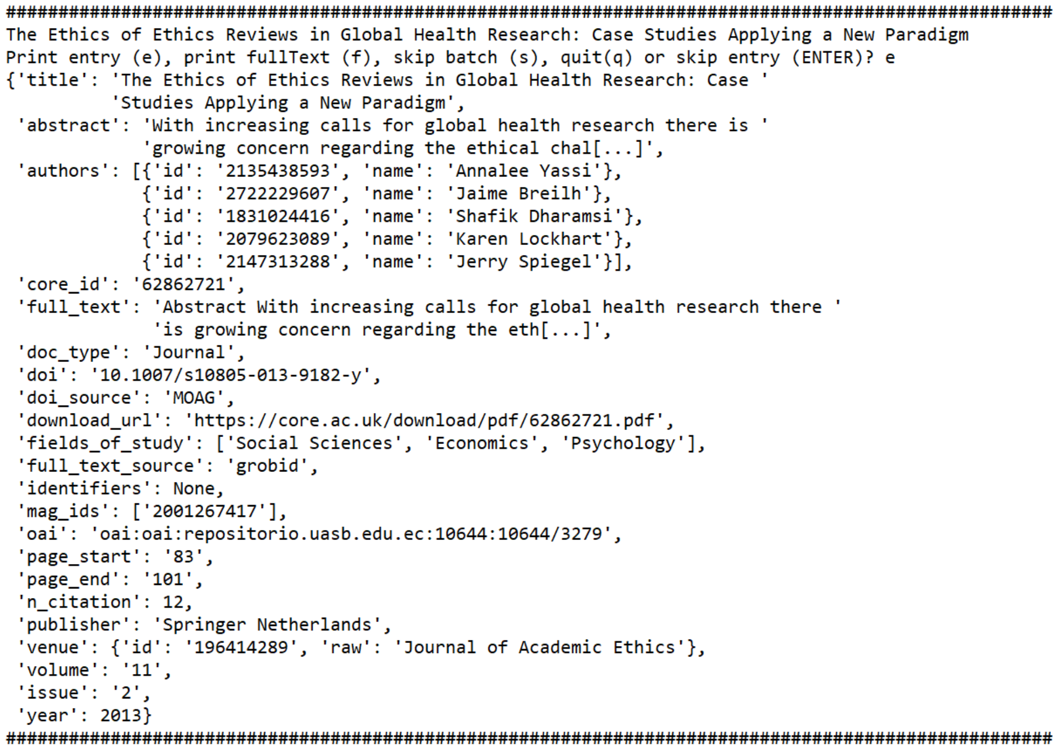
Putting together the corpus was mainly a challenge in terms of time- and resource efficiency. Three parts had to be put together: the CORE-data, the MAG-data, and some improved fulltexts resulting from other work in the webis-group. Those fulltexts replaced the partially flawed CORE-fulltexts for approximately 1M entries. Holding the relatively small MAG-data in memory was not an issue for this task, but both the CORE-data and the additionally extracted fulltexts could not be fully loaded at the same time. For the merging process, the CORE-data was processed line by line from file, while iteratively an ordered bundle of the next full texts needed was loaded into memory to avoid accessing and searching the respective files for every single entry. The following keys were extracted from the respective JSON lines and put together. Datapoints taken from MAG are: "doi", "magId", "title", "authors", "venue", "year", "n_citation", "page_start", "page_end", "doc_type", "venue", "volume", "issue", "publisher" and "fields_of_study" Datapoints taken from CORE are: "doi", "coreId", "abstract", "fullText", "oai", "identifiers" and "enrichments" which includes "citationCount" and "publisher" Possible sources of conflict between the data are the DOI, the year, citation count, and publisher name. Additionally, for 46K CORE entries, more than one match could be found in the MAG. In these cases, the information from the multiple matches was combined in the resulting entry, so possibly missing information like the start- or ending page could be put together from several matches. Conflicts in the sense of differing information between the entries occurred only in 86K cases and only related to the DOIs. In those cases, the MAG DOI from the first matching MAG entry was chosen and the CORE-DOI was deleted. This decision resulted from a control of 200 entries where CORE and MAG had differing DOIs for the same article via the

crossref.org API. In only 2 of the 200 cases, the CORE DOI proved to be correct when the MAG-DOI was not, while in 5 cases, neither of the two DOIs was correct. After these results, to always choose the MAG-DOI when in doubt seemed logical. (see process_2/doi_control.ipynb in the GitLab repository)
For 250K entries, the MAG entries contained double authors. To avoid including these, the mergescript checks authorlists for authors which appear more than once and only keeps the first entry of those authors for the final corpus. The script performing the merging can be found in the GitLab repository under mergescript.py.

4.3 Final Dataset

The merged corpus is stored in the form of files that each contain 100,000 entries. Each entry on its own is a valid JSON, containing the following keys.

Figure 4.1: JSON structure of the final corpus entries



```
####################################################################################################
The Ethics of Ethics Reviews in Global Health Research: Case Studies Applying a New Paradigm
Print entry (e), print fullText (f), skip batch (s), quit(q) or skip entry (ENTER)? e
{'title': 'The Ethics of Ethics Reviews in Global Health Research: Case '
          'Studies Applying a New Paradigm',
 'abstract': 'With increasing calls for global health research there is '
             'growing concern regarding the ethical chal[...]',
 'authors': [{'id': '2135438593', 'name': 'Annalee Yassi'},
             {'id': '2722229607', 'name': 'Jaime Breilh'},
             {'id': '1831024416', 'name': 'Shafik Dharamsi'},
             {'id': '2079623089', 'name': 'Karen Lockhart'},
             {'id': '2147313288', 'name': 'Jerry Spiegel'}],
 'core_id': '62862721',
 'full_text': 'Abstract With increasing calls for global health research there '
              'is growing concern regarding the eth[...]',
 'doc_type': 'Journal',
 'doi': '10.1007/s10805-013-9182-y',
 'doi_source': 'MOAG',
 'download_url': 'https://core.ac.uk/download/pdf/62862721.pdf',
 'fields_of_study': ['Social Sciences', 'Economics', 'Psychology'],
 'full_text_source': 'grobid',
 'identifiers': None,
 'mag_ids': ['2001267417'],
 'oai': 'oai:oai:repositorio.uasb.edu.ec:10644:10644/3279',
 'page_start': '83',
 'page_end': '101',
 'n_citation': 12,
 'publisher': 'Springer Netherlands',
 'venue': {'id': '196414289', 'raw': 'Journal of Academic Ethics'},
 'volume': '11',
 'issue': '2',
 'year': 2013}
####################################################################################################
```

All keys will be present, and have either an empty list or a None-value in those cases where the information wasn't included in the source data. Additionally to the keys from the sources, 'doi_source' and 'full_text_source' are

added. The first two specify the origin of the 'doi' and 'full_text'-Keys, the last refers to the classification of the fields of study according to the DFG-classification which assigns a number to each field. Details of the resulting corpus and properties of the included texts will be discussed in the following chapter.
The fulltexts in the corpus have not been preprocessed further, even though some include incorrect character encodings or large sections of data, tables and metadata. The decision to keep this raw data was made to avoid irreversibly cutting information from the full texts accidentally which may be relevant eventually. For the texts in their given form, some simple preprocessing steps should generate a human readable full text which can be used for authorship analysis purposes. The steps are identical to those taken in the section for full text sort out above, the module utils.py provides a corresponding method with the name "preprocessor" performing exactly those steps.
To browse through the JSON files of the corpus, the extraction_reader.py script in the github repository can be used. It allows to view whole entries and full texts of the corpus, and also has an option to automatically preprocess the full texts.

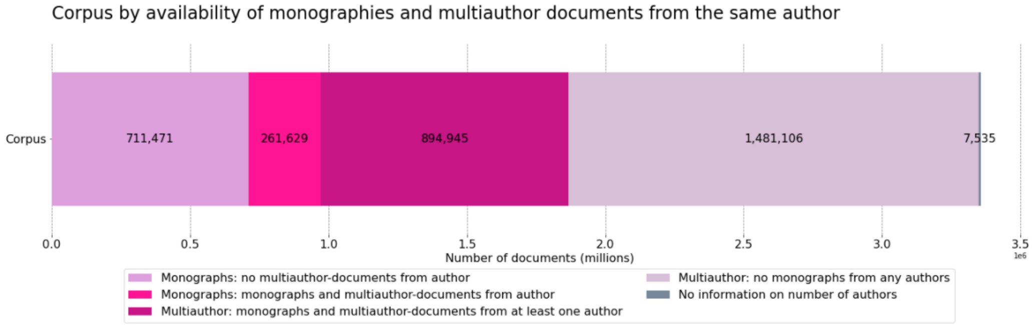
Chapter 5

Corpus Description

5.1 Corpus Size and Availability of Monographies and Multiauthor Documents

All figures and statistics presented in this chapter were generated via the notebook found in the GitLab repository under process_2/corpus_description.ipynb. The final corpus contains a total of 3,356,686 entries, each representing a scientific article with fulltext and metadata. A main purpose of the corpus is to make the comparison of multiauthor-documents and monographies by the same authors possible. The following Figure 5.1 illustrates the composition of the corpus regarding monographies and multiauthor-documents and the availability of documents by the same author.

Figure 5.1: Size of corpus by monographies and multiauthor documents



This means for 1,156,574 entries a comparison of monographies and multiauthor documents is possible. An additional 711,471 monographies for whose authors no multiauthor-documents are available can be used for the construction of authorship verification cases and for analysis of scientific writing styles.

For 1,481,106 multiauthor-documents, no monographies are available for any of the authors. For these texts, it will not be possible (or at least connected to the unproven assumption that the writing style could be attributed to the main author in large parts) to attribute writing styles to any of the listed authors. However, they can be useful in research cases where a large collection of texts from a certain research field is required, for example for general comparisons between subcorpora of humanities and sciences. They may also prove helpful in other contexts like the exploration of text reuse between scientific papers, independent of known authorship of reused passages. This led to the decision to include those documents in the corpus as well, even though they may not be ideally suited for the purpose of authorship verification tasks. The same applies for a small number of several thousand documents for which no author information at all is available.

5.2 Text Length

The corpus contains a broad range of texts of different lengths, which range from book-size entries to very short articles of just a few pages. The following metrics have been calculated on the texts without any preprocessing, so they reflect the length of the raw text entries, that also includes figures and appendixes.

Table 5.1: Number of documents in the corpus by textlength (characters)

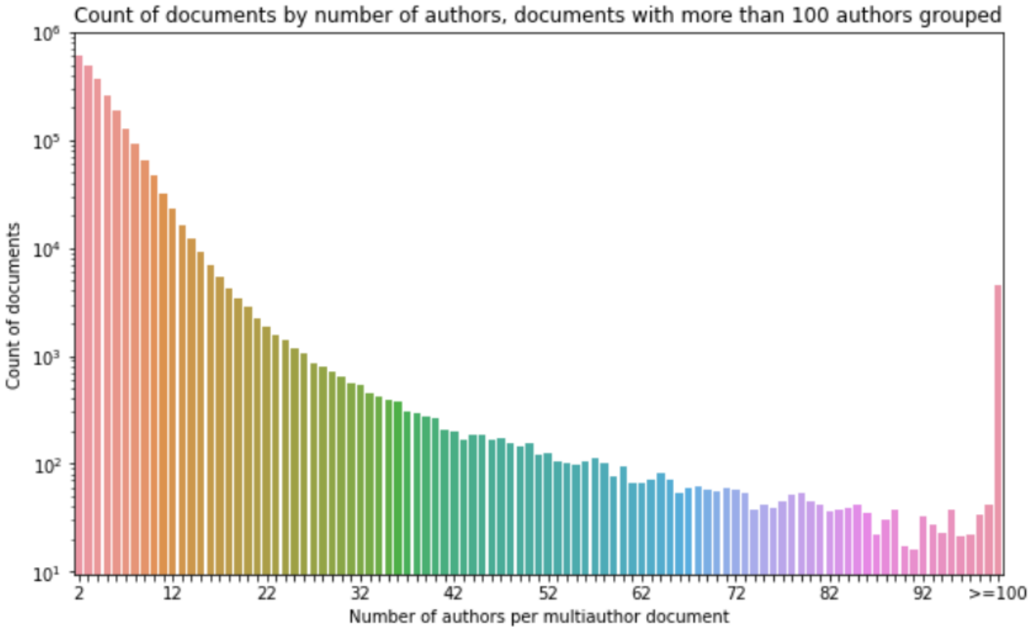
Number of documents by textlength (characters) of unprocessed full texts

Text length (characters)	Documents (total)	Monographies	Multiauthor documents
<= 3,000	39,300	13,680	25,567
3,001-5,000	96,067	32,059	63,382
5,001-50,000	2,273,246	467,844	1,799,435
50,001-250,000	771,756	301,975	468,473
> 250,000	176,317	157,542	18,744

As Table 5.1 shows, the largest part of the corpus has a length between 5,000 and 50,000 characters, which means a length of between 2 and 20 pages. Due to the size of the corpus, the number of documents with a length of 250,000 characters and more is still quite significant. It also becomes visible that the share of monographies among the long texts is rather high. Those seem to be mainly individual dissertations and less often collaborative book publications, which matches the impressions after a manual lookup. For more detailed text length statistics, see Table A.1 and Table A.2.

5.3 Number of Authors per Document

Figure 5.2: Number of documents by number of authors per document



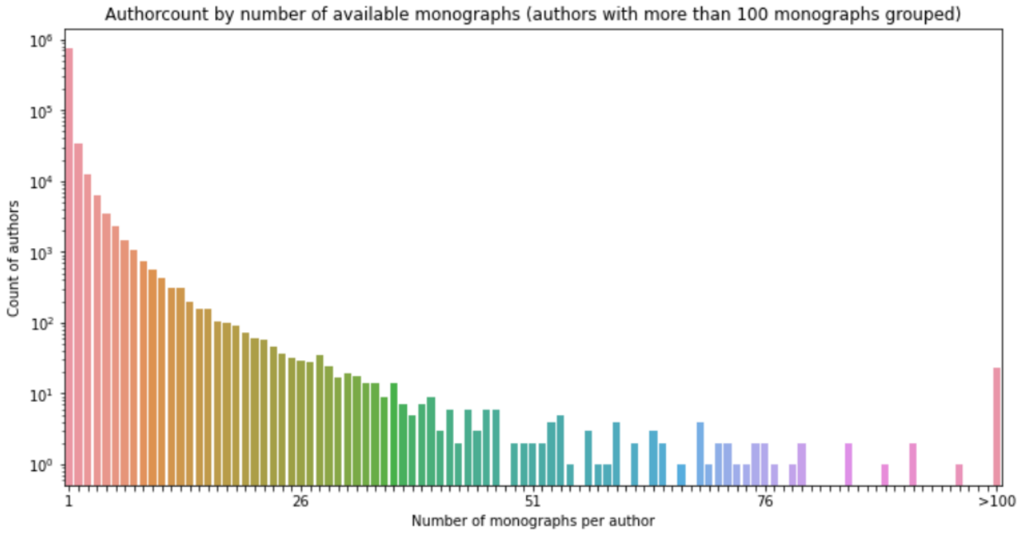
Multiauthor documents in the corpus have a surprising range of number of authors. While the total number of documents with more than two or three authors rapidly drop, a still high number of documents has 100 or more authors listed. The length of authorlists in the corpus goes as high as 5083 authors for a single paper. For a complete bar diagram of document-counts see Figure A.1. Papers with a very high number of authors are often the result of research in highly complex quantum or particle physics, the texts with the longest authorlists are all related to work with the ATLAS experiment at the particle collider of CERN and often contain an almost identical set of people.

5.4 Number of Monographies per Author

The 973,100 monographies in the corpus can be attributed to 795K different authors. Unsurprisingly, the largest part of those is represented with just one monography (92.02%). This also means, that 8% of the authors have written more than 24% of the monographies in the corpus. For authorship analysis, those will be especially relevant as a larger number of text by the author will certainly be able to improve results of authorship analysis methods.

Figure 5.3 illustrates the number of authors who have written a certain number of monographies in the corpus.

Figure 5.3: Number of available monographies per author

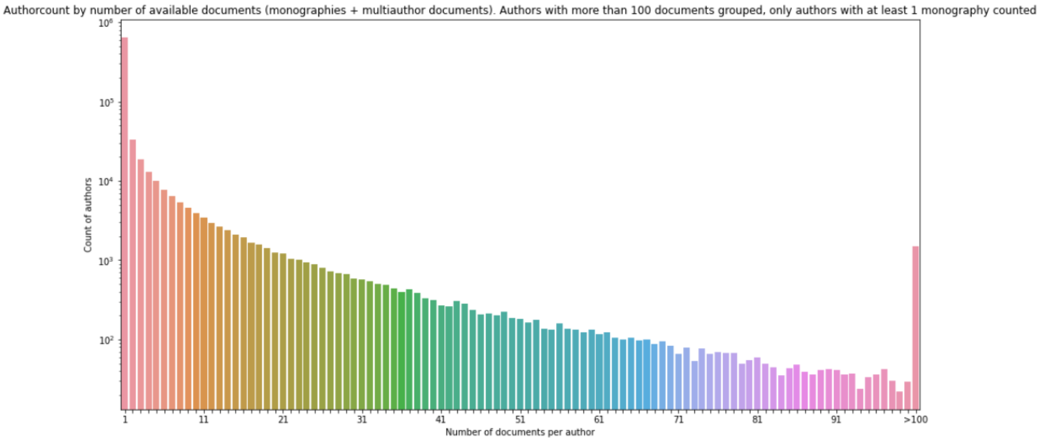


After all, most authors have written just a small number of monographies, but for few cases also numbers up to 602 monographies are possible. 8.6K authors have written more than 5, and still 2.5K have even written more than 10 monographies. Looking at the total number of documents (monographies and multiauthor-documents combined) available per author, the numbers in Figure 5.4 result. Authors without monographies are not considered in these numbers.

The by far largest part of authors are represented by just 1 or slightly more documents, but added up, for 70K authors more than 5 and for 42K authors more than 10 documents can be found in the corpus. The maximum number of available documents per author is 1092. Still, This means that the corpus allows the extraction of subsets that can be specified to contain a large number of texts by every single author, for specific use cases the number of available documents per author might even be above 100 and still contain a significant number of texts. However, the number of documents per author alone may not be the most relevant factor. If many multiauthor-documents, but only 1 or 2 monographies are available for an author, it may be difficult to identify this authors writing style with high confidence. This means, not only the overall number of documents per author, but also the combination of monographies and multiauthor documents is relevant.

For monography and multiauthor document counts without grouping of high

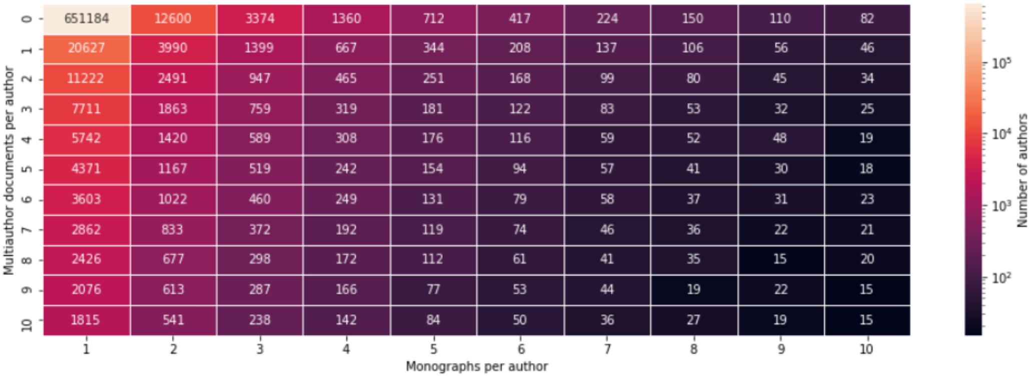
Figure 5.4: Number of available documents per author in total (authors with at least 1 monography only)



numbers, see Figure A.9 and Figure A.10.

5.5 Combination of Monographies and Multiauthor Documents

Figure 5.5: Number of authors by available monographies and multiauthor documents



Most authors in the corpus have written a small number of monographies and an even smaller number of multiauthor documents. The quick decline of the number of authors along each axis becomes very visible in Figure 5.5. As it can be expected, the by far largest part of the corpus are documents from authors with few monographs and few multiauthor documents. Looking at documents from authors with e.g. at least 3 monographies, numbers drop by several orders of magnitude compared to authors with only 1 monography. For smaller scale experimental settings, each cell of the above matrix still contains a number of documents that can be useful.

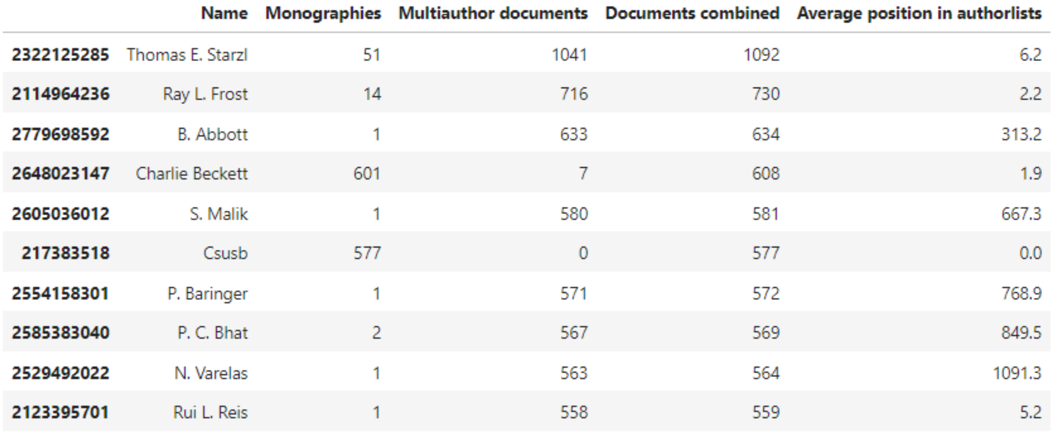
125.395 authors have written both monographies and multiauthor-documents in this corpus. They contributed to a total of 1.15M documents. When looking at the most frequent authors in the corpus, it becomes obvious that most of them have made contributions to multiauthor documents mainly. Taking the average position in the authorlist into consideration, it becomes more clear which authors may or may not have written any part of the texts. For example for P. Baringer, who is listed in position 845 of the author lists of his documents of average, it is rather unlikely to find many text really written by him among those 570 multiauthor documents he contributed to, as visible in Figure 5.6.

Looking at all of the 10 authors with the highest number of documents in the corpus, big differences in the number of documents they contributed to in prominent positions become clear. (see Figure 5.6) While the first three (Thomas Starzl, Ray Frost, B. Abbott) have several hundred of documents they contributed to in the first 5 positions of the authorlist, the following , if they have any multiauthor documents at all, have no more than 10 such documents. However, each of them has at least one multiauthor document in the corpus he is listed in as primary author, except for N. Varelas whose highest position in an authorlist is 3. This shows that the number of documents per author alone may not be the most relevant aspect for the potential of the corpus to trace the writing style of this author. A combination of the number of documents per author and the author position must be taken into consideration here.

5.6 Number of Authors per Document

As the number of authors that were involved in the writing of a document increases, the contribution of each of these authors will be getting smaller. The corpus contains a number of documents with a very high number of authors. Those are the result of large research processes where lots of small, earlier contribution led to the final paper. In terms of writing style, the by far largest share of these authors will not have contributed a single word or phrase to the

Figure 5.6: The 10 most frequent authors in the corpus, with number of available documents and average position in multiauthor-lists



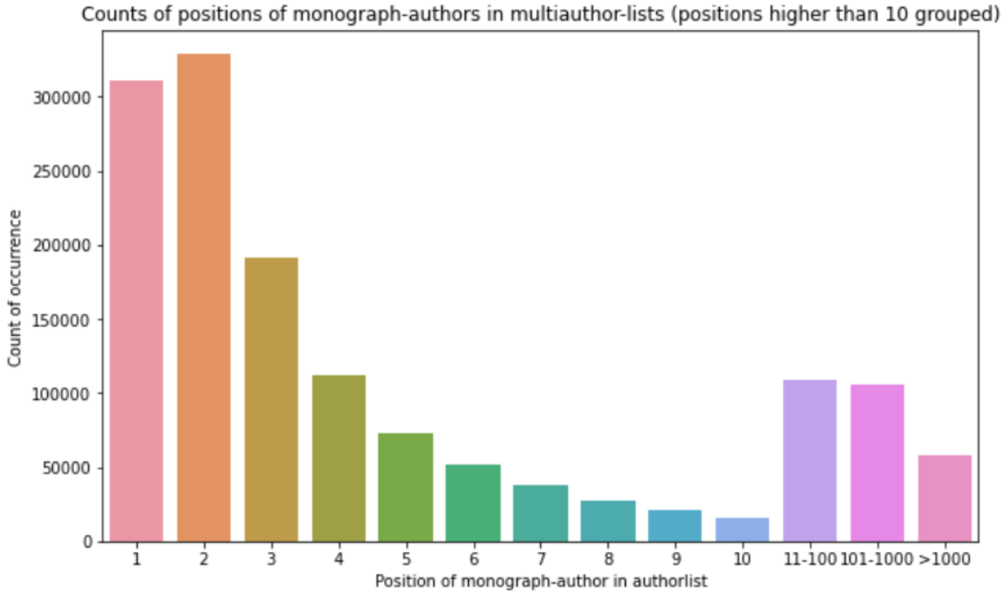
	Name	Monographies	Multiauthor documents	Documents combined	Average position in authorlists
2322125285	Thomas E. Starzl	51	1041	1092	6.2
2114964236	Ray L. Frost	14	716	730	2.2
2779698592	B. Abbott	1	633	634	313.2
2648023147	Charlie Beckett	601	7	608	1.9
2605036012	S. Malik	1	580	581	667.3
217383518	Csusb	577	0	577	0.0
2554158301	P. Baringer	1	571	572	768.9
2585383040	P. C. Bhat	2	567	569	849.5
2529492022	N. Varelas	1	563	564	1091.3
2123395701	Rui L. Reis	1	558	559	5.2

final text. It might make sense for many experiments, to limit the number of authors involved in the paper to avoid attributing a lot of multiauthor-documents to one person in which they have only been involved as position 20 or even 100 in the long author-list. Examples have been discussed earlier in this chapter as seen in Figure 5.6.
Publications with very large authorlists don't constitute a large part of the corpus. However, they disproportionally influence the number of documents available per author, as demonstrated in the cases above. This should be kept in mind when considering the numbers of documents per author in the two preceding chapters.

Figure 5.7 shows the accumulated total frequency of positions which monography authors take in the authorlists of their multiauthor documents.

While most frequently listed in the first two positions, a significant share is also placed in positions higher than 10. This must be taken into consideration for further work with the corpus. In many cases, it might make sense to consider a high author position as equal to the author having made no contribution at all, and to include only the first authors as people who have written the paper. This will be discussed further in chapter 6 regarding useful selection criteria for subcorpora and introducing a feature to leave out high author positions when collecting all documents by certain authors. A figure without grouped bars for high positions can be found in Figure A.8.

Figure 5.7: Position of monography-authors in multiauthor-lists (positions higher than 10 grouped)



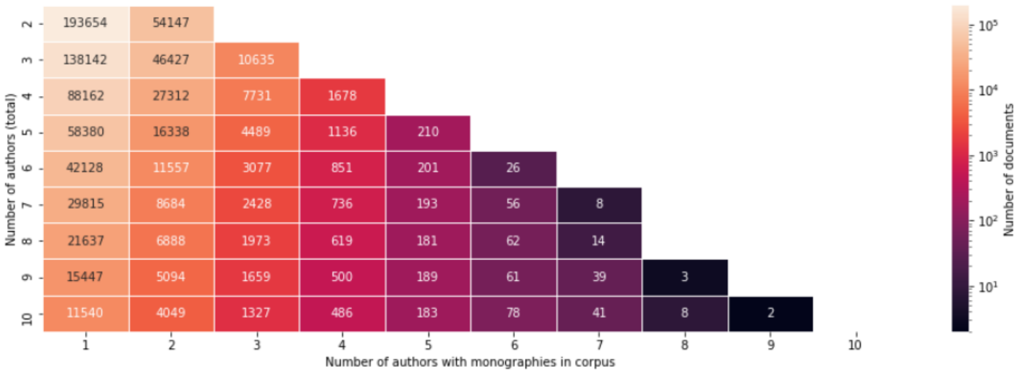
5.7 Multiauthor Documents and their Share of Authors with Monographies

Not only the number of texts available for a certain author may be an important selection criterium. For multiauthor documents, the share of authors with available documents is of particular interest. If for a multiauthor document a matching monography is available for only one out of ten authors, the probability to find the writing style of this specific person in the final paper decreases significantly. If not by chance this one person is the main author, it may be hard to verify any hypothesis on the authorship of the document. Even if an authorship verification method attributes the text to the one monography author with high confidence, only a limited number of hypotheses on the authorship of the document can be tested at all, as it is impossible to control its similarity to the style of the remaining 9 authors for which no individual writing samples are available.

If monographies are known for all of the authors, a document can be analysed regarding the individual writing styles of all of them. In this case, probability of finding traces of the individuals is highly increased, and it might even become possible to attribute certain passages of the paper to some of its individual authors. Figure 5.8 illustrates the number of multiauthor-documents with up to 10 authors in total and the share of authors with available monographies among those.

5.8 Combination of Monographies and Multiauthor Documents

Figure 5.8: Documents by number of authors and number of authors with available monographies



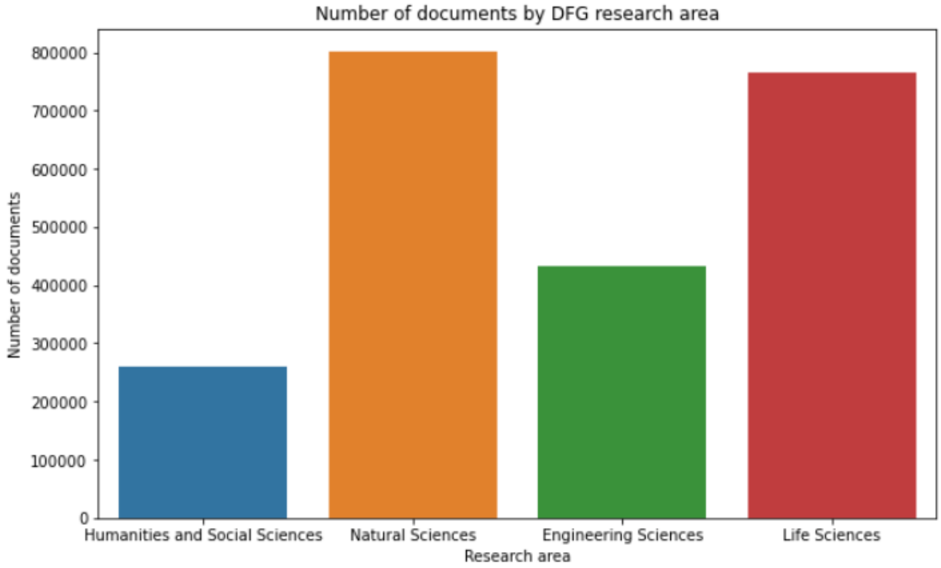
Again, the number of documents steeply decreases the higher the count on both axis gets. For documents with 2, 3 or 4 authors, many documents can be found that allow to make some analyses with monographies of all of the authors. For 1.3M of the multiauthor documents, at least half of the authors have at least one monography included in the corpus. For 66K documents, all of the authors have also a monography. Regarding subcorpora, this allows the construction of smaller sets where all or most of the multiauthors have monographies, or for large sets where only 1 or 2 of the authors must have individual monographies.

5.9 Academic Disciplines

The corpus contains fields of study information for 1.7M of its entries. The original fields of study from the MOAG have been mapped on the review board structure of the German Research Association to make them more uniform and give them a standardized hierarchical structure. [Deutsche Forschungsgemeinschaft, 2021] The corpus contains documents from 46 distinct review boards.

Not included are only ‘Chemical solid state and surface research’, while ‘Physical Chemistry’, ‘Analytical Chemistry’ and ‘Theoretical Chemistry’ have been collected under one label together as they could not be sufficiently precisely separated from each other. The total number of documents for all fields can be found in Figure A.2.

Figure 5.9: Number of documents by scientific area



The DFG system classifies the research boards in 4 main research areas. Figure 5.9 shows that Humanities and Engineering Sciences are less frequent than Natural and Life Sciences, but for all areas a sufficient number of texts can be found in the corpus. Some differences between academic practices and writing styles become directly visible when looking at basic corpus statistics.

While 29% of the documents in the corpus are monographies, this share is smaller when looking at the documents with fields of studies. Fields information seem to be available for multiauthor documents more frequently. However, the share of monographies varies between the areas – in the humanities, 22% of the available documents are monographies, in Life Sciences, this share is just above 6%. This also reflects in the median number of authors per document, which is significantly higher in Life Sciences. While e.g. Philosophy, Fine Arts and History have a median number of authors per document of just 1, this number is 6 for documents in Virology and Immunology.

It is also not surprising that texts from the realms of the humanities are often longer, with a median value which is 40% higher than for Natural Sciences. Electrical Engineering and Information Technology are only a median of 19K

Table 5.2: Monographies, multiauthor documents, authors per document and textlength by research area

Statistics by research area				
Area	Monographies	Multiauthor documents	Authors per document (median)	Textlength in characters (median)
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

characters long, while philosophical papers have a median of 45K characters. (for more detailed statistics, see Figure A.4) The corpus seems to reflect initial assumptions about differences in the writing style of different disciplines, especially between humanities and the three other areas.

5.10 Summary of Corpus Statistics

As shown, the large amount of documents in the corpus allows to select texts by a large number of criteria, which each on their own are still met by a significant number of texts. Selecting only very short texts is equally possible as selecting book-size texts or choosing to include only authors with a high number of monographies and will still generate subsets with several thousands of texts. Even selecting only texts for which individual writing style analyses are possible on all the authors leaves a subset of more than 70,000 documents. If subsets of just a few hundred entries are sufficient for a specific experiment, it is also possible to combine many of these constraints and to for example look at multiauthor texts from a range of humanities only, for which all authors have a monography on their own as well. After all, one of the advantages of the large corpus is that it contains subsets tailored for a broad range of research questions. The goal of the final chapter of this thesis is to introduce a simple tool that allows the selection of texts according to the described criteria.

Chapter 6

Corpus Explorer

6.1 A Database for the Corpus

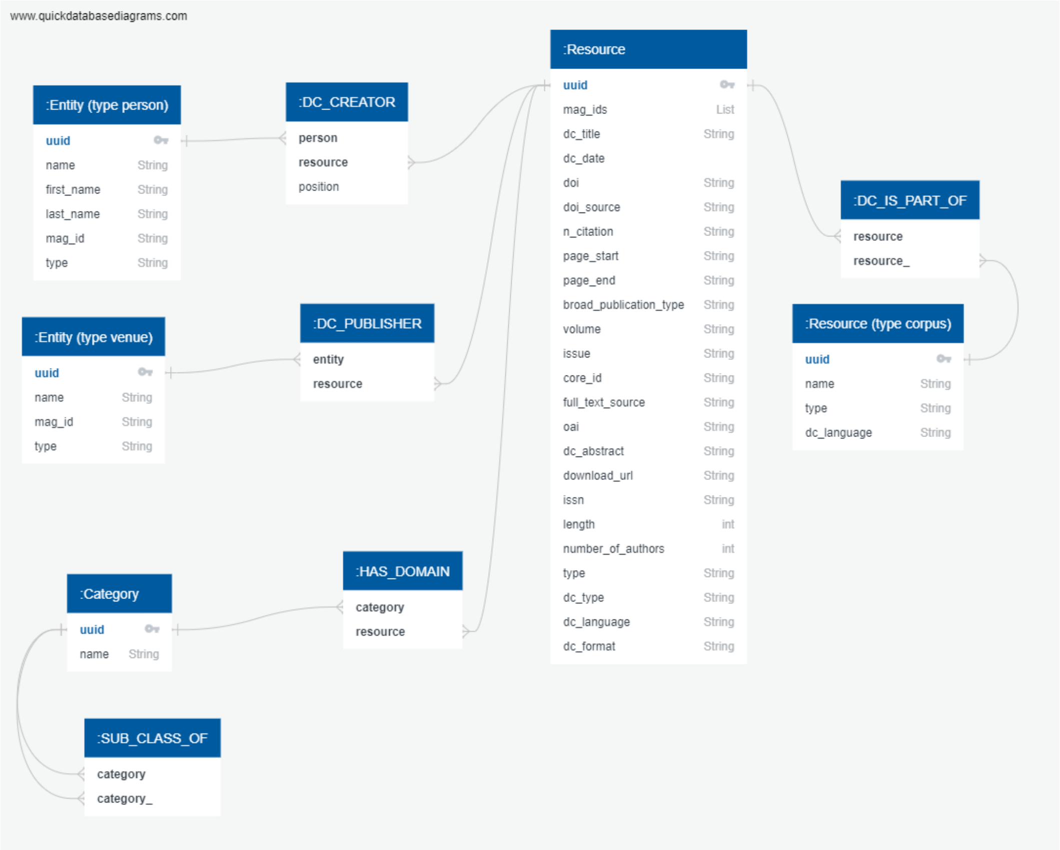
To efficiently explore, query and extract the data from the corpus, a fitting data-structure had to be found. In this case, the corpus-data is integrated in the webis-groups Peak Authorship neo4J database. Then, a browser interface was developed that allows connecting to the database, make queries on it and export the search results to a certain directory. Neo4J is a native graph database. It is schema optional and stores its data in the form of nodes and pointers to other nodes with which a specific relation is established. It uses its own query language called Cypher.

6.2 Data Schema

As stated above, neo4J is schema optional, but for clearly structured data establishing a schema made sense. The schema is formulated according to the specifications of the Dublin Core Metadata Initiative. [DCMI Usage Board, 20.01.2020] The DCMI provides a number of documents and standards for the description of documents. In the DCMI Metadata Terms, a standardized vocabulary for metadata of texts is specified, which will be used for the description of texts and their relations in the corpus graph.
The graph contains four main types of nodes: :Resource, :Person, :Entitiy and :Category.

1. Resources
   Three kinds of resources are specified via their type-property, which can be corpus or document. All document nodes are related to one basic corpus node via a :DC_IS_PART_OF relationship, which can be used

Figure 6.1: Graph Schema



to retrieve all documents belonging to the corpus. Resource nodes contain most of the document information also included in the JSON-lines described before. Only the full texts are excluded, to limit the size of the nodes at a reasonable level. The full list of included attributes are:
mag_ids, dc_title, dc_date, publisher, doi, doi_source,
n_citation, page_start, page_end, publication_type,
volume, issue, core_id, full_text_source, fields_of_study,
oai, identifiers, abstract, download_url, issn, length,
number_of_authors, uuid, type, dc_type,
dc_language, dc_format
dc_date contains only the year of publication. The dc_type of all resources is 'text' and broadly classifies the resource, while the PeakAuthorship-specific type references whether a resource is a corpus, a document or some other type of collection. Additionally, publication type may contain information whether a resource is a dissertation, a presentation, a paper

or a book. The quite similar names of these properties are not ideal, but derive from the specific nature of the data and the terminology of the DCMI.

2. Person
   Person nodes only contain the basic name, surname and last name of an author, as well as their uuid and MAG-Id. They reference natural individuals having written a text. They are linked to the documents they contributed to via a :DC_CREATOR relation, which has as attribute the position of the respective person in the authorlist.

3. Entity
   Other than authors, entities reference the institutions or publishing companies or venues responsible for the publication of the documents. As attributes, they contain only their uuid, MAG-Id and their name. They are linked to the documents they published via a :DC_PUBLISHER relation. This relation is used for individual editors or publishing institutions according to the DCMI, but will be only for the latter in our graph.

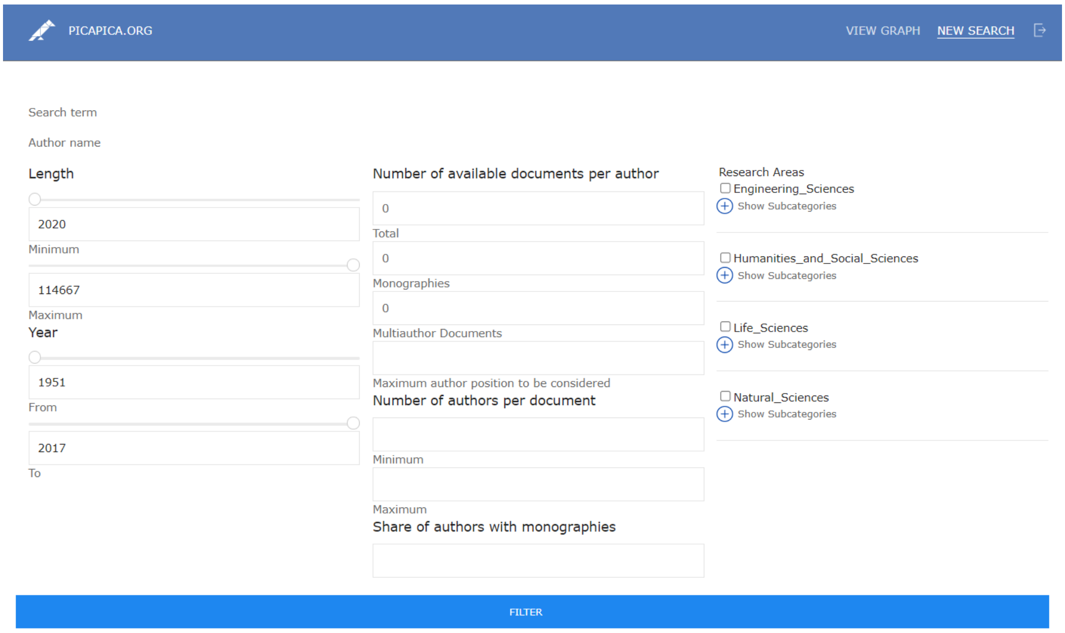
4. Category
   Category nodes are present on two levels. They can be the individual review boards of the DFG classification, or the broader research areas those review boards belong to. Both contain their name and a uuid. Documents are connected to their review boards via a :HAS_GENRE relation. Review boards are connected to their research area via a :SUB_CLASS_OF relation. Both relation types are not part of the DCMI terminology, but are specifics of the PeakAuthorship database. Research areas have no direct connection to any resources, and can only be linked to those via the review boards.

To access the full texts of the documents, the respective entries have to be extracted separately from the files specified in chapter 4. This does not allow to browse through the fulltexts via the database, but makes faster querying possible. The abstracts of the documents will be included to give a quick insight into the contents. All datapoints included in the nodes and the type of relationships used are displayed in Figure 6.1.

6.3 Interface

The corpus interface itself is integrated in the PicaPica corpus explorer browser interface. It consists of four relevant endpoints – one start page that shows an initial visualization of a small part of the graph, one search endpoint that allows to enter the desired features into a search mask, (see Figure 6.2) one results-endpoint that lists the entries and allows to go though them page by page, (see Figure 6.3) and a view_graph endpoint that visualizes search results. (see Figure 6.4)
The interface is implemented using Flask. To be able to start, the application has to be able to connect to the database via credentials provided in a settings-file, otherwise execution will abort.

Figure 6.2: Screenshot of search interface



Search term

Author name

Length

2020
Minimum

114667
Maximum

Year

1951
From

2017
To

Number of available documents per author

0
Total

0
Monographies

0
Multiauthor Documents

Maximum author position to be considered

Number of authors per document

Minimum

Maximum

Share of authors with monographies

Research Areas
□ Engineering_Sciences
Show Subcategories

□ Humanities_and_Social_Sciences
Show Subcategories

□ Life_Sciences
Show Subcategories

□ Natural_Sciences
Show Subcategories

The search function (Figure 6.2) includes the following selection-features:

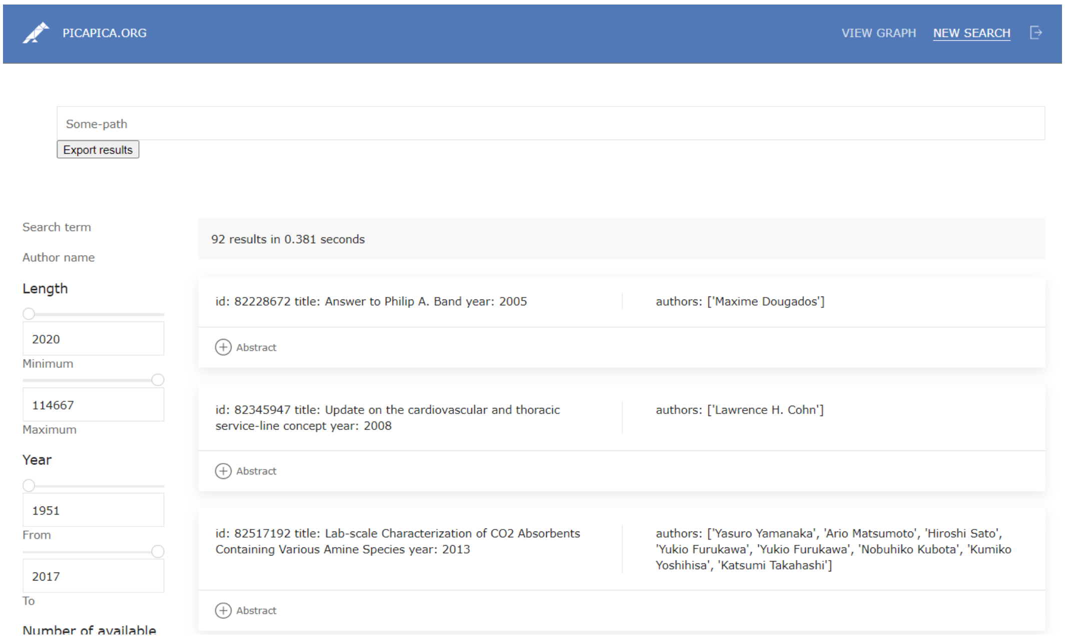
1. Text length (minimum and maximum)
2. Publication year (minimum and maximum)
3. The minimum number of monographies, multiauthor documents and documents in total to be available for each included author
4. The minimum and maximum number of authors per document

5. The Fields of study the documents are linked with

6. A share of authors for which monographies are available in the corpus (to find only documents for which monographies are available by all authors, the user would have to enter 1.0 here)

7. The maximum position of an author in the authorlist of the respective document to still be considered. This feature practically alows to "cut" authorlists at a certain position to avoid getting results for a certain author where this specific person is only listed as 50th or 100th author. It also allows to search only for primary or secondary authors, if stronger restrictions are required.

Figure 6.3: Screenshot: a page with search results



Some-path

Search term

Author name

Length

2020
Minimum

114667
Maximum

Year

1951
From

2017
To

Number of available

92 results in 0.381 seconds

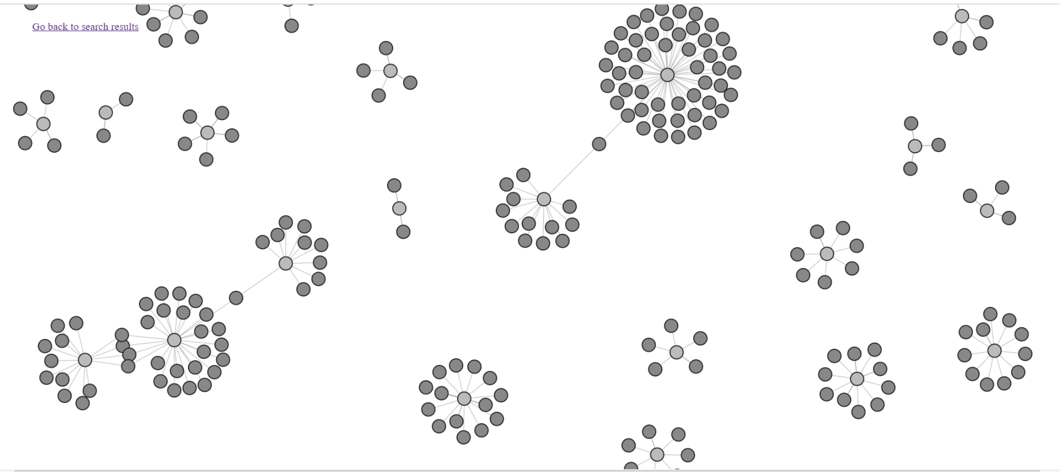
id: 82228672 title: Answer to Philip A. Band year: 2005 | authors: ['Maxime Dougados']

id: 82345947 title: Update on the cardiovascular and thoracic service-line concept year: 2008 | authors: ['Lawrence H. Cohn']

id: 82517192 title: Lab-scale Characterization of CO2 Absorbents Containing Various Amine Species year: 2013 | authors: ['Yasuro Yamanaka', 'Ario Matsumoto', 'Hiroshi Sato', 'Yukio Furukawa', 'Yukio Furukawa', 'Nobuhiko Kubota', 'Kumiko Yoshihisa', 'Katsumi Takahashi']

According to the search features specified, the program will put together a cypher query and issue this to the database. The results returned will be serialized and displayed in the results endpoint page-wise. (Figure 6.3) The user can the go through the results and look at basic properties of the results, like the title, authorlist and year of publication. Individual abstracts can be read in a toggle menu to keep the result list as a whole clear and organized. The results endpoint also has an option to export the search results to a specific directory. For this, not the database nodes, but the JSON entries from the complete corpus files will be extracted. As the corpus files are quite large, exporting entries may take some time and should not be expected to be complete within mere minutes.

Figure 6.4: Screenshot: graph visualization



Additionally, the search results can be visualized in a seperate browser window. (Figure 6.4) The visualization uses the D3.js library and works efficiently for few hundreds of nodes. For queries with a very high number of results, it may suffer from performance difficulties. Its main purpose should be displaying the connection of documents from a handful of authors, or to explore the connectivity of small clusters of highly connected documents or of individual authors with a relatively high number of documents in the corpus.

Using the corpus explorer requires the installation of the python packages specified in the requirements file of the respective directory. Afterwards, executing the main.py program will start the flask interface. The user has to provide the access information to the Neo4J database in the settings.py file to allow the program to connect to the database. The complete source code for the interface can be found in the GitLab repository under interface/src, where additional information for usage can be found in the readme.md. At the moment of submission of this thesis, the dataset is not yet available in the webis Neo4J database due to some difficulties with the database settings, in particular the database's import basepath which at the moment does not allow the database to access the dataset directories. As soon as those difficulties are resolved, the complete dataset will be uploaded and available in the authorship.test database. Until then, testing the explorer will require to upload some test data to a local database. A suitable test dataset in CSV form and an indexer which allows to upload the CSVs to a local databse can be found in the GitLab repoistory under data/test-csvs.

Chapter 7

Conclusion

The goal of this thesis was to explore the potential of the CORE dataset for the construction of a corpus of scientific texts for research on authorship analysis. The great strength of CORE definetly lies in its large number of full texts of open access research, problems of the dataset mainly in its incomplete language tagging and the missing possibility to identify authors beyond name identities. Combining the CORE dataset with data from the Microsoft Academic Graph was able to match a high number of entries between both datasets and allowed to fill gaps of one dataset with data from the other. The Core dataset mainly provided full texts and PDF links, while the graph structure of the MOAG could be used to easily identify documents written by the same author and link a high number of documents with their respective research fields.

The large intersection between both datasets allowed to include more than 3.5 million documents in the corpus. Checking those for processable English full-texts and excluding documents with a high portion of text extraction errors and unrecognizable text still left a total of 3,356,745 documents. 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.

This makes the SAAD significantly larger than corpora in previous academic work, like e.g. the one used by Ammar et al. [2018], and, even more importantly, thematically significantly more diverse. The corpus includes metadata linking the documents to their respective fields of study for around half of its entries. Those fields of study, derived from the MOAG, have been mapped onto the research classification of the German Research Association and cover all fields of this classification. Even for rather niche areas, at least several hundred entries can be found in the corpus. Concerning the 4 main research areas of the classification, all are represented by at least 250,000 documents (in the case of the Humanities). This means, the corpus allows to differentiate

between academic motivations and methods on a sufficiently specific level, and to discriminate between Humanities and Natural Sciences, but also between medicinal and psychological, or economic and historical research. It also offers a broad range regarding the length of its texts, their period of publication and the required number of documents by specific author.
Over all, the requirements for the corpus formulated in chapter 1 are mostly fulfilled. The corpus combines monographies and multiauthor documents for 123.564 different authors which contributed to 1,144,915 documents. Texts have been checked for their readability and correct language tagging. Linking the documents to a scientific discipline was possible for just about half of the documents, so this is the only requirement which could only be satisfied partially.

Finally, the tool presented in chapter 6 should allow to explore the corpus and allow basic selection of subcorpora fulfilling criteria based on the specific research interests of the user. It also should allow dealing with a main problem of the corpus: the partially high number of authors in some documents, which make textual contributions of the authors listed in high positions highly unlikely.
Future research with the corpus could include the application of well tested authorship analysis methods like Unmasking 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.

Appendix A

Appendix

A.1 Source Code

All code used in this thesis can be found under `https://git.webis.de/code-teaching/theses/thesis-sauer/`.

A.2 Additional Figures and Tables

Table A.1: Number of documents by textlength (words)

Number of documents by textlength (words) of unprocessed full texts	
Text length(words)	Documents (total)
From 0 to 500	57,596
From 501 to 1,000	148,303
From 1,001 to 10,000	2,463,792
From 10,001 to 40,000	513,298
More than 40,001	173,697

Figure A.1: Number of documents by number of authors per document

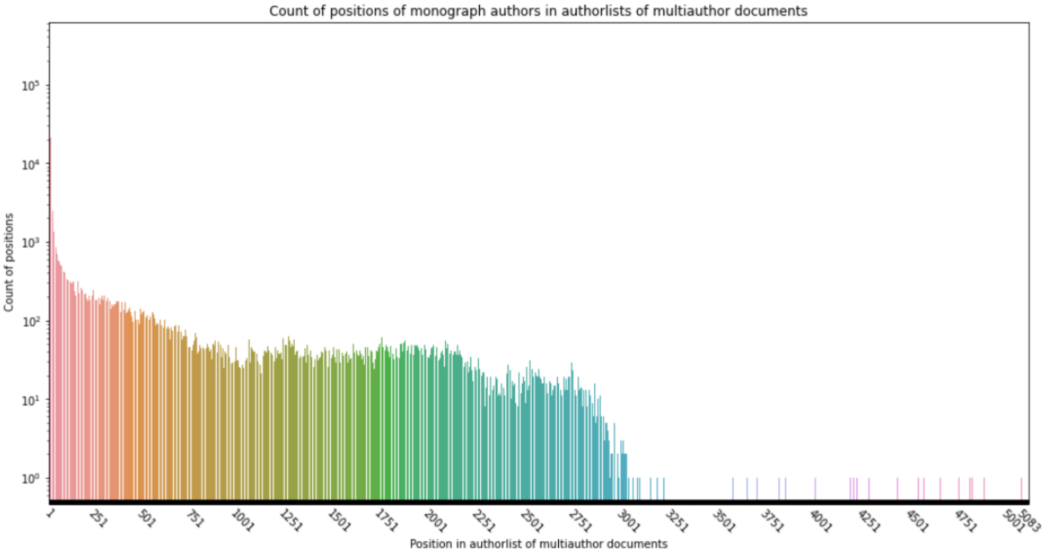
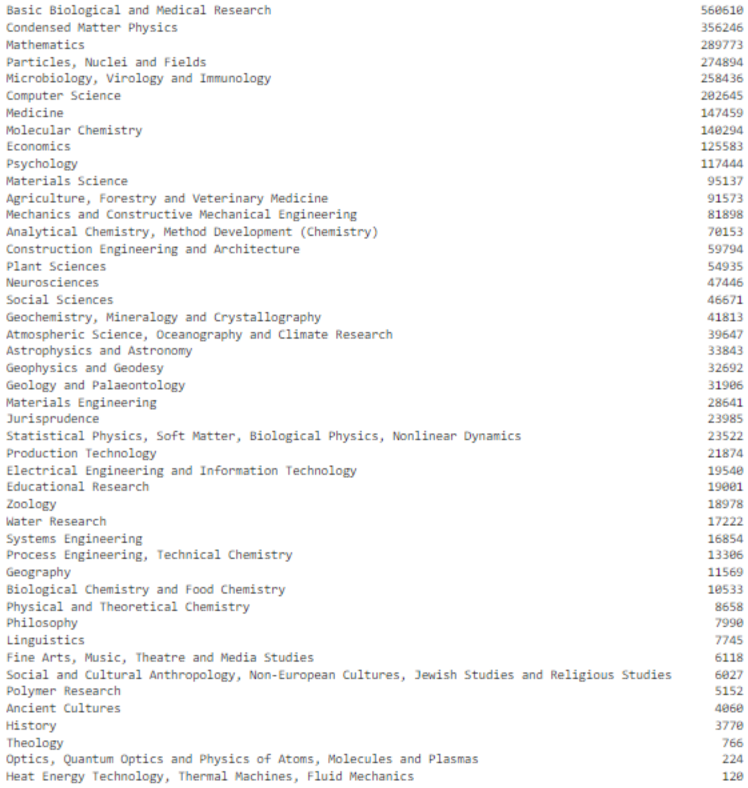
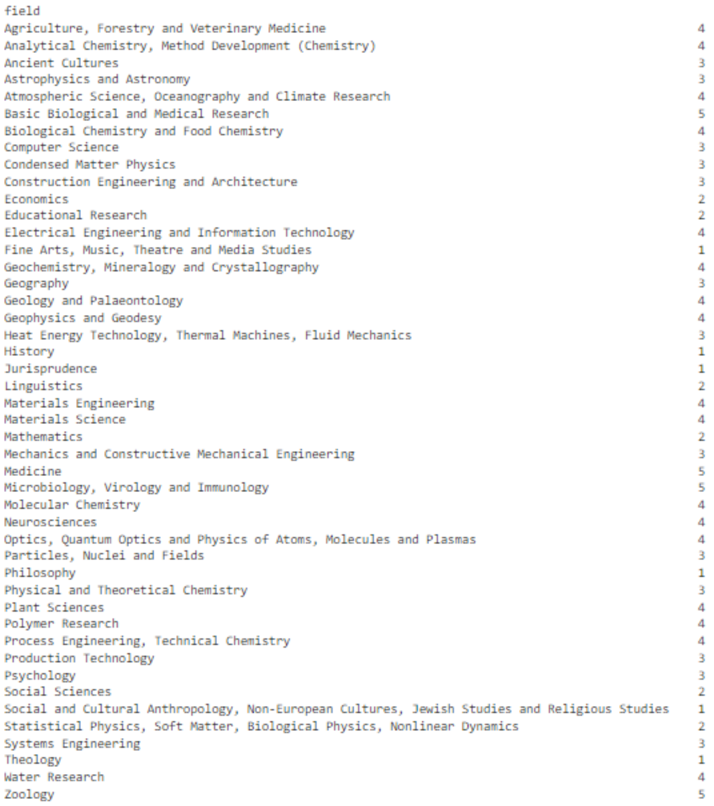


Figure A.2: Number of documents by fields of study



Basic Biological and Medical Research	560610
Condensed Matter Physics	356246
Mathematics	289773
Particles, Nuclei and Fields	274894
Microbiology, Virology and Immunology	258436
Computer Science	202645
Medicine	147459
Molecular Chemistry	140294
Economics	125583
Psychology	117444
Materials Science	95137
Agriculture, Forestry and Veterinary Medicine	91573
Mechanics and Constructive Mechanical Engineering	81898
Analytical Chemistry, Method Development (Chemistry)	70153
Construction Engineering and Architecture	59794
Plant Sciences	54935
Neurosciences	47446
Social Sciences	46671
Geochemistry, Mineralogy and Crystallography	41813
Atmospheric Science, Oceanography and Climate Research	39647
Astrophysics and Astronomy	33843
Geophysics and Geodesy	32692
Geology and Palaeontology	31906
Materials Engineering	28641
Jurisprudence	23985
Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics	23522
Production Technology	21874
Electrical Engineering and Information Technology	19540
Educational Research	19001
Zoology	18978
Water Research	17222
Systems Engineering	16854
Process Engineering, Technical Chemistry	13306
Geography	11569
Biological Chemistry and Food Chemistry	10533
Physical and Theoretical Chemistry	8658
Philosophy	7990
Linguistics	7745
Fine Arts, Music, Theatre and Media Studies	6118
Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies	6027
Polymer Research	5152
Ancient Cultures	4060
History	3770
Theology	766
Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas	224
Heat Energy Technology, Thermal Machines, Fluid Mechanics	120

Figure A.3: Median of authors per document by fields of study

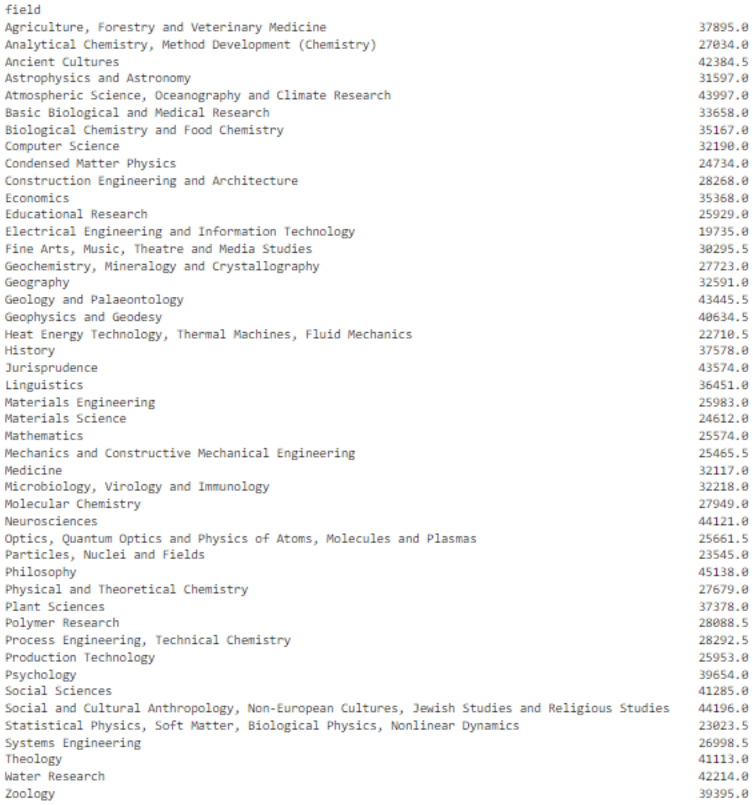


```
field
Agriculture, Forestry and Veterinary Medicine                                                  4
Analytical Chemistry, Method Development (Chemistry)                                           4
Ancient Cultures                                                                               3
Astrophysics and Astronomy                                                                     3
Atmospheric Science, Oceanography and Climate Research                                         4
Basic Biological and Medical Research                                                          5
Biological Chemistry and Food Chemistry                                                        4
Computer Science                                                                               3
Condensed Matter Physics                                                                       3
Construction Engineering and Architecture                                                      3
Economics                                                                                      2
Educational Research                                                                           2
Electrical Engineering and Information Technology                                              4
Fine Arts, Music, Theatre and Media Studies                                                    1
Geochemistry, Mineralogy and Crystallography                                                   4
Geography                                                                                      3
Geology and Palaeontology                                                                      4
Geophysics and Geodesy                                                                         4
Heat Energy Technology, Thermal Machines, Fluid Mechanics                                      3
History                                                                                        1
Jurisprudence                                                                                  1
Linguistics                                                                                    2
Materials Engineering                                                                          4
Materials Science                                                                              4
Mathematics                                                                                    2
Mechanics and Constructive Mechanical Engineering                                              3
Medicine                                                                                       5
Microbiology, Virology and Immunology                                                          5
Molecular Chemistry                                                                            4
Neurosciences                                                                                  4
Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas                             4
Particles, Nuclei and Fields                                                                   3
Philosophy                                                                                     1
Physical and Theoretical Chemistry                                                             3
Plant Sciences                                                                                 4
Polymer Research                                                                               4
Process Engineering, Technical Chemistry                                                       4
Production Technology                                                                          3
Psychology                                                                                     3
Social Sciences                                                                                2
Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies  1
Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics                       2
Systems Engineering                                                                            3
Theology                                                                                       1
Water Research                                                                                 4
Zoology                                                                                        5
```

Table A.2: Mean and median lextlengths

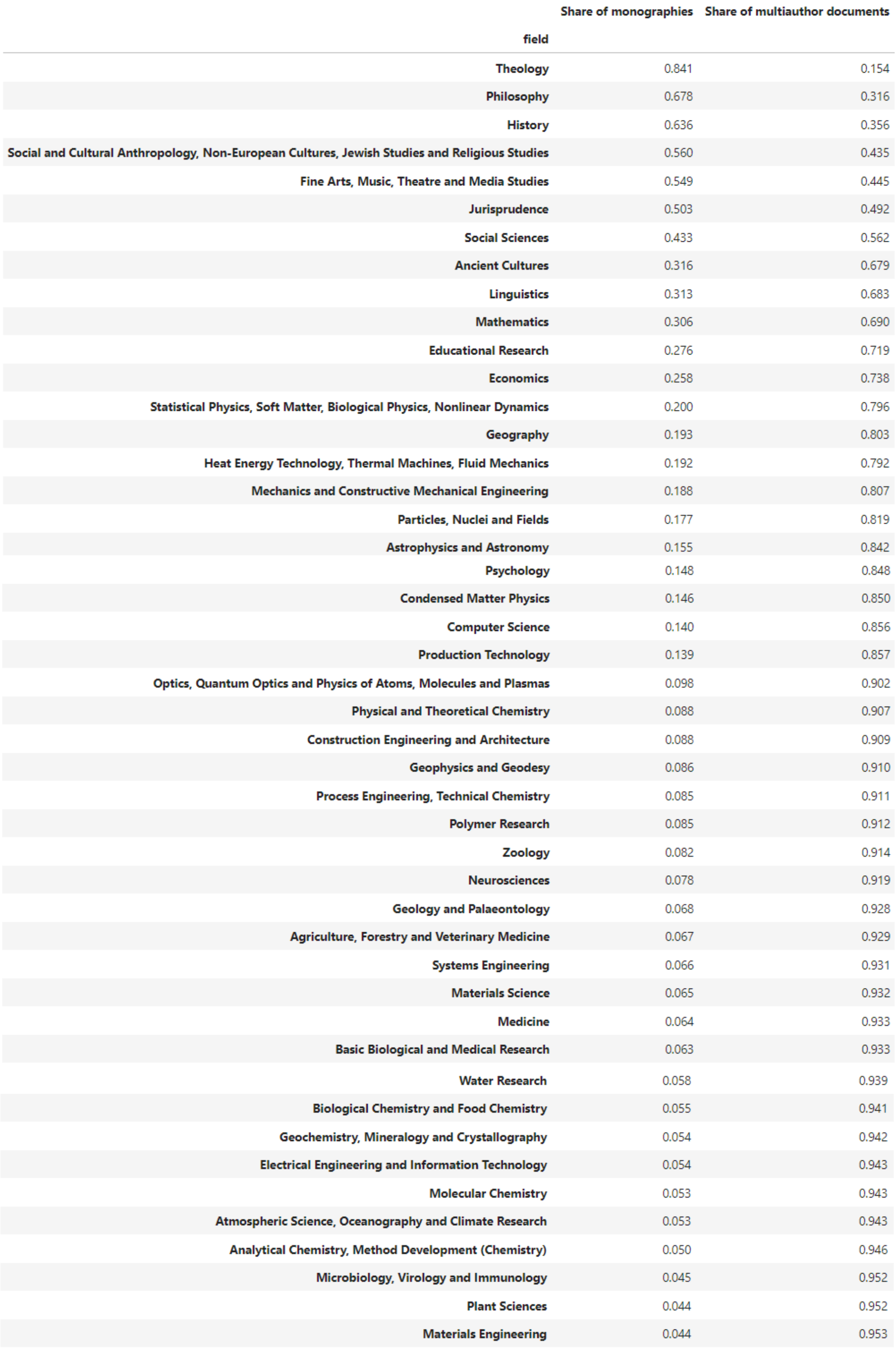
Characters	Mean textlength (all)	62,854
	Median textlength (all)	31,234
	Mean textlength (monographies)	122,086
	Median textlength (monographies)	44864
	Mean textlength (multiauthor)	38,686
	Median textlength (multiauthor)	29,071
Words	Mean textlength (all)	10,026
	Median textlength (all)	5,008

Figure A.4: Median document length in characters by fields of study



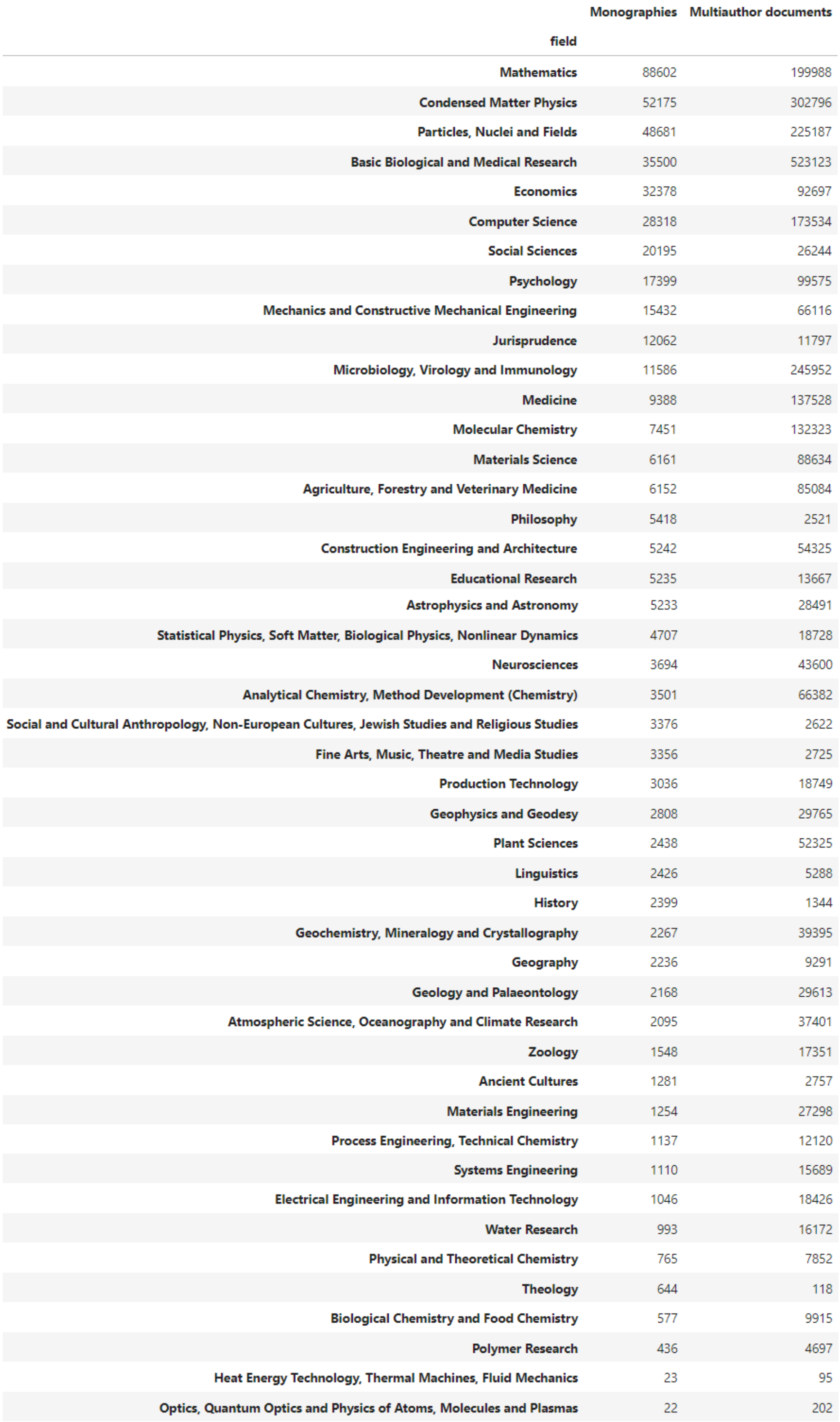
```
field
Agriculture, Forestry and Veterinary Medicine                                                    37895.0
Analytical Chemistry, Method Development (Chemistry)                                             27034.0
Ancient Cultures                                                                                 42384.5
Astrophysics and Astronomy                                                                       31597.0
Atmospheric Science, Oceanography and Climate Research                                           43997.0
Basic Biological and Medical Research                                                            33658.0
Biological Chemistry and Food Chemistry                                                          35167.0
Computer Science                                                                                 32190.0
Condensed Matter Physics                                                                         24734.0
Construction Engineering and Architecture                                                        28268.0
Economics                                                                                        35368.0
Educational Research                                                                             25929.0
Electrical Engineering and Information Technology                                                19735.0
Fine Arts, Music, Theatre and Media Studies                                                      30295.5
Geochemistry, Mineralogy and Crystallography                                                     27723.0
Geography                                                                                        32591.0
Geology and Palaeontology                                                                        43445.5
Geophysics and Geodesy                                                                           40634.5
Heat Energy Technology, Thermal Machines, Fluid Mechanics                                        22710.5
History                                                                                          37578.0
Jurisprudence                                                                                    43574.0
Linguistics                                                                                      36451.0
Materials Engineering                                                                            25983.0
Materials Science                                                                                24612.0
Mathematics                                                                                      25574.0
Mechanics and Constructive Mechanical Engineering                                                25465.5
Medicine                                                                                         32117.0
Microbiology, Virology and Immunology                                                            32218.0
Molecular Chemistry                                                                              27949.0
Neurosciences                                                                                    44121.0
Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas                               25661.5
Particles, Nuclei and Fields                                                                     23545.0
Philosophy                                                                                       45138.0
Physical and Theoretical Chemistry                                                               27679.0
Plant Sciences                                                                                   37378.0
Polymer Research                                                                                 28088.5
Process Engineering, Technical Chemistry                                                         28292.5
Production Technology                                                                            25953.0
Psychology                                                                                       39654.0
Social Sciences                                                                                  41285.0
Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies    44196.0
Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics                         23023.5
Systems Engineering                                                                              26998.5
Theology                                                                                         41113.0
Water Research                                                                                   42214.0
Zoology                                                                                          39395.0
```

Figure A.5: Percentage of monographies and multiauthor documents by fields of study



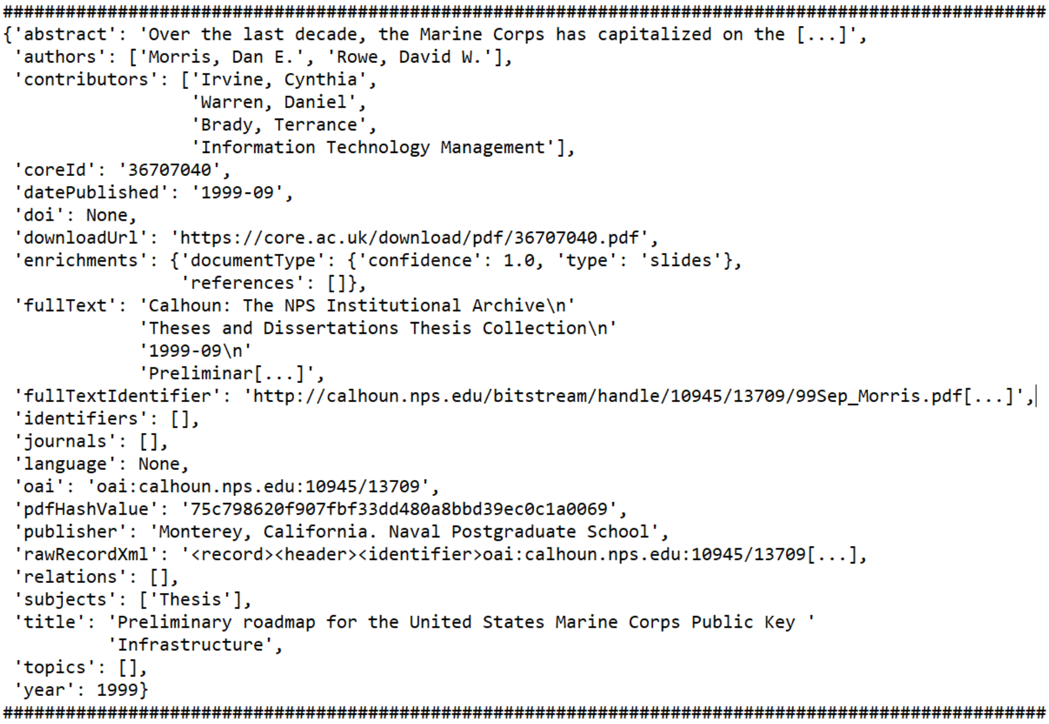
	Share of monographies	Share of multiauthor documents
field		
Theology	0.841	0.154
Philosophy	0.678	0.316
History	0.636	0.356
Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies	0.560	0.435
Fine Arts, Music, Theatre and Media Studies	0.549	0.445
Jurisprudence	0.503	0.492
Social Sciences	0.433	0.562
Ancient Cultures	0.316	0.679
Linguistics	0.313	0.683
Mathematics	0.306	0.690
Educational Research	0.276	0.719
Economics	0.258	0.738
Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics	0.200	0.796
Geography	0.193	0.803
Heat Energy Technology, Thermal Machines, Fluid Mechanics	0.192	0.792
Mechanics and Constructive Mechanical Engineering	0.188	0.807
Particles, Nuclei and Fields	0.177	0.819
Astrophysics and Astronomy	0.155	0.842
Psychology	0.148	0.848
Condensed Matter Physics	0.146	0.850
Computer Science	0.140	0.856
Production Technology	0.139	0.857
Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas	0.098	0.902
Physical and Theoretical Chemistry	0.088	0.907
Construction Engineering and Architecture	0.088	0.909
Geophysics and Geodesy	0.086	0.910
Process Engineering, Technical Chemistry	0.085	0.911
Polymer Research	0.085	0.912
Zoology	0.082	0.914
Neurosciences	0.078	0.919
Geology and Palaeontology	0.068	0.928
Agriculture, Forestry and Veterinary Medicine	0.067	0.929
Systems Engineering	0.066	0.931
Materials Science	0.065	0.932
Medicine	0.064	0.933
Basic Biological and Medical Research	0.063	0.933
Water Research	0.058	0.939
Biological Chemistry and Food Chemistry	0.055	0.941
Geochemistry, Mineralogy and Crystallography	0.054	0.942
Electrical Engineering and Information Technology	0.054	0.943
Molecular Chemistry	0.053	0.943
Atmospheric Science, Oceanography and Climate Research	0.053	0.943
Analytical Chemistry, Method Development (Chemistry)	0.050	0.946
Microbiology, Virology and Immunology	0.045	0.952
Plant Sciences	0.044	0.952
Materials Engineering	0.044	0.953

Figure A.6: Number of monographies and multiauthor documents by fields of study



	Monographies	Multiauthor documents
field		
Mathematics	88602	199988
Condensed Matter Physics	52175	302796
Particles, Nuclei and Fields	48681	225187
Basic Biological and Medical Research	35500	523123
Economics	32378	92697
Computer Science	28318	173534
Social Sciences	20195	26244
Psychology	17399	99575
Mechanics and Constructive Mechanical Engineering	15432	66116
Jurisprudence	12062	11797
Microbiology, Virology and Immunology	11586	245952
Medicine	9388	137528
Molecular Chemistry	7451	132323
Materials Science	6161	88634
Agriculture, Forestry and Veterinary Medicine	6152	85084
Philosophy	5418	2521
Construction Engineering and Architecture	5242	54325
Educational Research	5235	13667
Astrophysics and Astronomy	5233	28491
Statistical Physics, Soft Matter, Biological Physics, Nonlinear Dynamics	4707	18728
Neurosciences	3694	43600
Analytical Chemistry, Method Development (Chemistry)	3501	66382
Social and Cultural Anthropology, Non-European Cultures, Jewish Studies and Religious Studies	3376	2622
Fine Arts, Music, Theatre and Media Studies	3356	2725
Production Technology	3036	18749
Geophysics and Geodesy	2808	29765
Plant Sciences	2438	52325
Linguistics	2426	5288
History	2399	1344
Geochemistry, Mineralogy and Crystallography	2267	39395
Geography	2236	9291
Geology and Palaeontology	2168	29613
Atmospheric Science, Oceanography and Climate Research	2095	37401
Zoology	1548	17351
Ancient Cultures	1281	2757
Materials Engineering	1254	27298
Process Engineering, Technical Chemistry	1137	12120
Systems Engineering	1110	15689
Electrical Engineering and Information Technology	1046	18426
Water Research	993	16172
Physical and Theoretical Chemistry	765	7852
Theology	644	118
Biological Chemistry and Food Chemistry	577	9915
Polymer Research	436	4697
Heat Energy Technology, Thermal Machines, Fluid Mechanics	23	95
Optics, Quantum Optics and Physics of Atoms, Molecules and Plasmas	22	202

Figure A.7: Structure of CORE JSON-lines



```
##############################################################################################
{'abstract': 'Over the last decade, the Marine Corps has capitalized on the [...]',
 'authors': ['Morris, Dan E.', 'Rowe, David W.'],
 'contributors': ['Irvine, Cynthia',
                  'Warren, Daniel',
                  'Brady, Terrance',
                  'Information Technology Management'],
 'coreId': '36707040',
 'datePublished': '1999-09',
 'doi': None,
 'downloadUrl': 'https://core.ac.uk/download/pdf/36707040.pdf',
 'enrichments': {'documentType': {'confidence': 1.0, 'type': 'slides'},
                 'references': []},
 'fullText': 'Calhoun: The NPS Institutional Archive\n'
             'Theses and Dissertations Thesis Collection\n'
             '1999-09\n'
             'Preliminar[...]',
 'fullTextIdentifier': 'http://calhoun.nps.edu/bitstream/handle/10945/13709/99Sep_Morris.pdf[...]',
 'identifiers': [],
 'journals': [],
 'language': None,
 'oai': 'oai:calhoun.nps.edu:10945/13709',
 'pdfHashValue': '75c798620f907fbf33dd480a8bbd39ec0c1a0069',
 'publisher': 'Monterey, California. Naval Postgraduate School',
 'rawRecordXml': '<record><header><identifier>oai:calhoun.nps.edu:10945/13709[...],
 'relations': [],
 'subjects': ['Thesis'],
 'title': 'Preliminary roadmap for the United States Marine Corps Public Key '
          'Infrastructure',
 'topics': [],
 'year': 1999}
##############################################################################################
```

Figure A.8: Positions of monography-authors in authorlists of multiauthor-documents

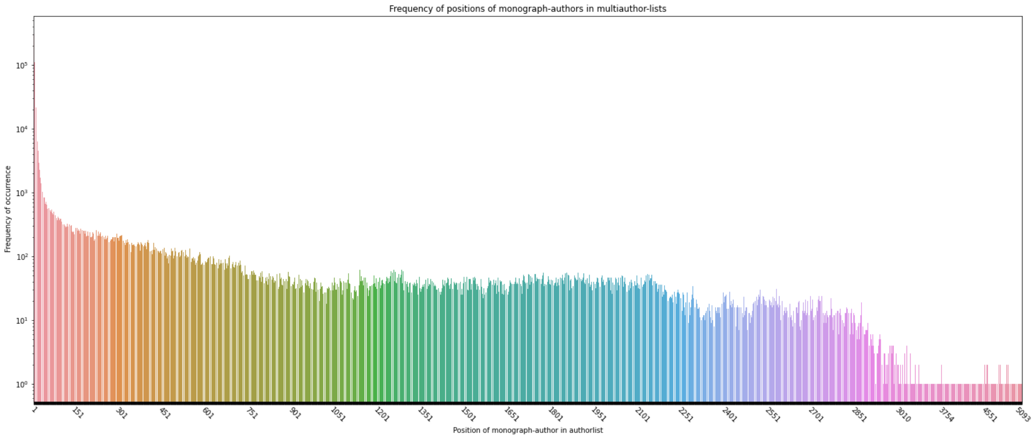


Figure A.9: Count of authors by number of monographies

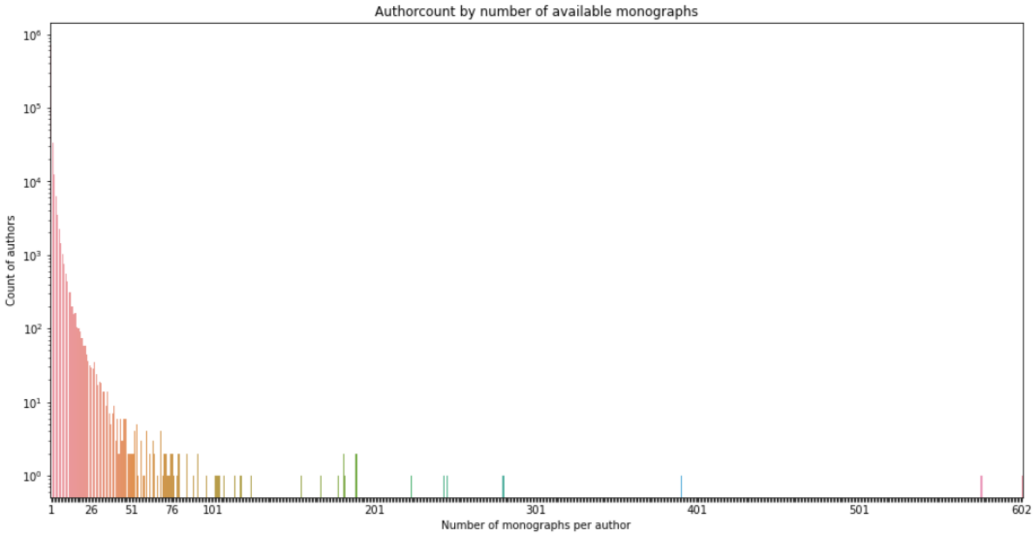
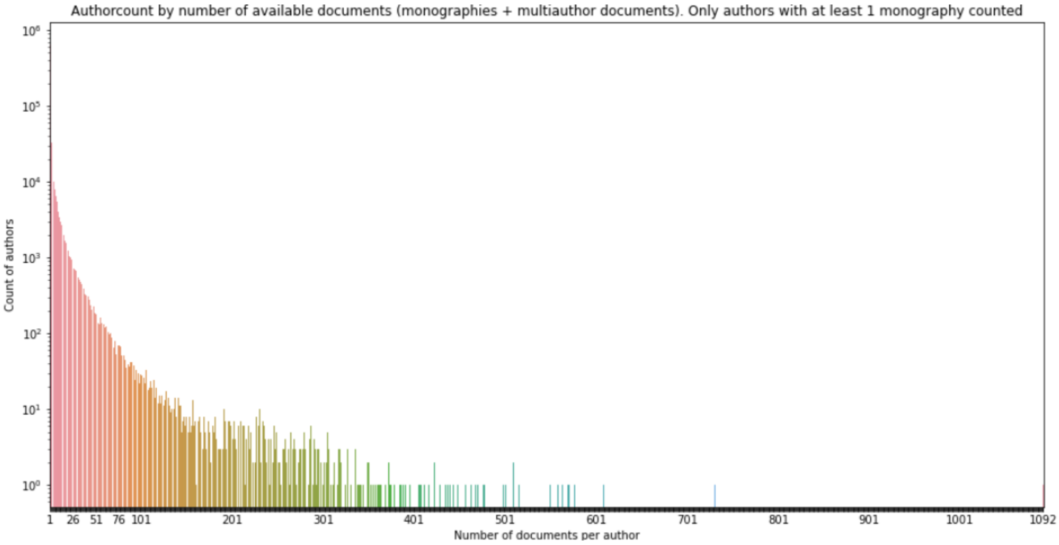


Figure A.10: Count of authors by number of documents (monographies and multiauthor documents combined). Only authors with at least 1 monography counted



Bibliography

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 Willhelm, Zheng Yuan, Madeleine Zuylen, and oren. Construction of the literature graph in semantic scholar. In Srinivas Bangalore, Jennifer Chu-Carroll, and Yunyao Li, editors, *Proceedings of the 2018 Conference of the North American Chapter of*, pages 84–91, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-3011. 2.3, 7

Alberto Barrón-Cedeño, Martin Potthast, Paolo Rosso, Benno Stein, and Andreas Eiselt. Corpus and evaluation measures for automatic plagiarism detection. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *7th Conference on International Language Resources and Evaluation (LREC 2010)*. European Language Resources Association (ELRA), 2010. ISBN 2-9517408-6-7. 2.2

Iz Beltagy, Lo Kyle, and Arman Cohan. Scibert: A pretrained language model for scientific text. *EMNLP*, 2019. URL `https://arxiv.org/pdf/1903.10676`. 2.3

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Generalizing unmasking for short texts. In Jill Burstein, Christie Doran, and Thamar Solorio, editors, *14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)*, pages 654–659. Association for Computational Linguistics, 2019a. URL `https://www.aclweb.org/anthology/N19-1068`. 2.1.1, 2.1.2

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Bias analysis and mitigation in the evaluation of authorship verification. In Anna

Korhonen, Lluís Màrquez, and David Traum, editors, *57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 6301–6306. Association for Computational Linguistics, 2019b. URL https://www.aclweb.org/anthology/P19-1634. 2.1.2

Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. Authorship verification for short messages using stylometry. In *International Conference on Computer, Information and Telecommunication Systems (CITS), 2013*, pages 1–6, Piscataway, NJ, 2013. IEEE. ISBN 978-1-4799-0168-5. doi: 10.1109/CITS.2013.6705711. 2.1.1

Daniel T. Citron and Paul Ginsparg. 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, 2015. doi: 10.1073/pnas.1415135111. 2.3

DCMI Usage Board. Dcmi metadata terms, 20.01.2020. URL https://dublincore.org/specifications/dublin-core/. 6.2

Deutsche Forschungsgemeinschaft. Review boards: Structure, 2021. URL https://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp. 5.9

Darrin Eide and Charles Huang. Microsoft academic graph data schema, 28.05.2021. URL https://docs.microsoft.com/en-us/academic-services/graph/reference-data-schema. 3.2

Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys*, 52(6): 1–42, 2020. ISSN 0360-0300. doi: 10.1145/3345317. 2.2

Bela Gipp, Norman Meuschke, and Corinna Breitinger. Citation-based plagiarism detection: Practicability on a large-scale scientific corpus. *Journal of the Association for Information Science and Technology*, 65(8):1527–1540, 2014. ISSN 23301635. doi: 10.1002/asi.23228. 2.2, 2.3

Lynne Horwood, Shirley Sullivan, Eve Young, and Jane Garner. Oai compliant institutional repositories and the role of library staff. *Library Management*, 25(4/5):170–176, 2004. ISSN 0143-5124. doi: 10.1108/01435120410533756. 3.1

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In *The 2020 Web Conference (WWW)*, 2020. URL https://www.microsoft.com/en-us/research/publication/heterogeneous-graph-transformer/. 3.2

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651 Titel anhand dieser ArXiv-ID in Citavi-Projekt übernehmen*, 2016a. 4.1.2

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016b. 4.1.2

Patrick Juola. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334, 2007. ISSN 1554-0669. doi: 10.1561/1500000005. 2.1.1

Mike Kestemont, Kim Luyckx, Walter Daelemans, and Thomas Crombez. Cross-genre authorship verification using unmasking. *English Studies*, 93 (3):340–356, 2012. ISSN 0013-838X. doi: 10.1080/0013838X.2012.668793. 2.1.2

Petr Knoth and Nancy Pontika. Aggregating research papers from publishers? systems to support text and data mining: Deliberate lack of interoperability or not? In Richard Eckart de Castilho, Sophia Ananiadou, Thomas Margoni, Wim Peters, and Stelios Piperidis, editors, *INTEROP2016*, 2016. URL `http://oro.open.ac.uk/46870/`. 3.1

Petr Knoth and Zdenek Zdrahal. Core: Three access levels to underpin open access. *D-Lib Magazine*, 18(11/12), 2012. doi: 10.1045/november2012-knoth. 3.1

Petr Knoth, Vojtech Robotka, and Zdenek Zdrahal. 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, 2011. URL `http://oro.open.ac.uk/32180/`. 3.1

Mirco Kocher and Jacques Savoy. A simple and efficient algorithm for authorship verification. *Journal of the Association for Information Science and Technology*, 68(1):259–269, 2017. ISSN 23301635. doi: 10.1002/asi.23648. 2.1.2

Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In Carla Brodley, editor, *Proceedings of the 41st annual Design Automation Conference*, page 62, New York, NY, 2004. ACM. ISBN 1581138285. doi: 10.1145/1015330.1015448. 1.1, 2.1.2

Moshe Koppel and Yaron Winter. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187, 2014. ISSN 23301635. doi: 10.1002/asi.22954. 2.1.2

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009. ISSN 15322882. doi: 10.1002/asi.20961. 1.1, 2.1.1

Frederick Mosteller and David L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275, 1963. ISSN 01621459. doi: 10.2307/2283270. 2.1.1

Heather Piwowar, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. The state of oa: a large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, 6:e4375, 2018. ISSN 2167-8359. doi: 10.7717/peerj.4375. 3.1

Nektaria Potha and Efstathios Stamatatos. Improved algorithms for extrinsic author verification. *Knowledge and Information Systems*, 62(5):1903–1921, 2020. ISSN 0219-1377. doi: 10.1007/s10115-019-01408-4. 2.1.2

Shachar Seidman. *Authorship verification using the impostors method: Notebook for PAN at CLEF 2013*. 2013. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.667.4579&rep=rep1&type=pdf`. 1.1, 2.1.2

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In Aldo Gangemi, Stefano Leonardi, and Alessandro Panconesi, editors, *Proceedings of the 24th International Conference on World Wide Web*, pages 243–246, New York, NY, USA, 05182015. ACM. ISBN 9781450334730. doi: 10.1145/2740908.2742839. 3.2

Felipe Soares, Viviane Pereira Moreira, and Karin Becker. A large parallel corpus of full-text scientific articles, 2019. URL `https://arxiv.org/pdf/1905.01852`. 2.3

Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009. ISSN 15322882. doi: 10.1002/asi.21001. 2.1.2

Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82, 2011. ISSN 1574-020X. doi: 10.1007/s10579-010-9115-y. 2.2

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7386–7393, 2019. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33017386. 2.3