A Large-Scale Analysis of Cross-Lingual Citations in English Papers

Tarek Saier and Michael Färber

Institute AIFB, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
{tarek.saier,michael.faerber}@kit.edu

Abstract Citation data is an important source of insight into the scholarly discourse and the reception of publications. Outcomes of citation analyses and the applicability of citation based machine learning approaches heavily depend on the completeness of citation data. One particular shortcoming of scholarly data nowadays is language coverage. That is, non-English publications are often not included in data sets, or language metadata is not available. While national citation indices exist, these are often not interconnected to other data sets. Because of this, citations between publications of differing languages (cross-lingual citations) have only been studied to a very limited degree. In this paper, we present an analysis of cross-lingual citations based on one million English papers, covering three scientific disciplines and a time span of 27 years. Our results unveil differences between languages and disciplines, show developments over time, and give insight into the impact of cross-lingual citations on scholarly data mining as well as the publications that contain them. To facilitate further analyses, we make our collected data and code for analysis publicly available.

Keywords: Scholarly Data, Citations, Cross-Lingual, Citation Analysis

1 Introduction

Citations are an essential tool for scientific practice. By allowing authors to refer to existing publications, citations make it possible to position one's work within the context of others', critique, compare, and point readers to supplementary reading material. In other words, citations enable scientific discourse. Because of this, citations are a valuable indicator for the academic community's reception of and interaction with published works. Their analysis is used, for example, to quantify research output [12], qualify references [1], and detect trends [2]. Furthermore, citations can be utilized to aid researchers through, for example, summarization [6] or recommendation [25, 7] of papers, and through applications driven by document embeddings in general [3].

Because such analyses and applications require data to be based on, the availability of citation data or lack thereof is decisive with regard to the areas, in which respective insights can be gained and approaches developed. Here, the literature points in two mayor directions with much potential for improvement—namely the humanities [4, 18] and non-English publications [32, 22, 26, 27]. Due

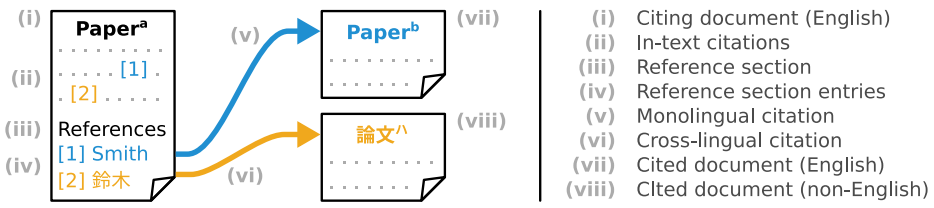

Figure 1 Schematic explanation of terminology.

to citation data's lack of language coverage and lack of language metadata, a particular practice not well researched so far is cross-lingual citation. That is, references where the citing and cited documents are written in different languages (see *(vi)* in Figure 1). Because English is currently the de facto academic lingua franca, citations from non-English languages to English can be assumed to generally be significantly more prevalent than the other way around. This dichotomy is reflected in existing literature, where usually either citations from English [18, 21], or to English [31, 15, 16, 29] are analyzed. As both directions involve a non-English document on one side of the citation, the analysis of either is challenging with today's Anglocentric state of citation data.

To add to the body of work studying cross-lingual citations *from English*, we perform a large-scale analysis on one million documents and address the following research questions.[1]

1. How prevalent are English to non-English references? We consider prevalence in general, in different disciplines, across time, and within publications that use them.
2. Is self-citation a driving factor for citing non-English work?
3. Are non-English works deemed "citable" in the context of English papers?
4. Do cross-lingual citations pose a particular challenge for data mining?
5. Does citing other languages impact the success of a publication?

Through our analysis, we make the following contributions.

1. We give insight into cross-lingual citations in English papers at a scale, that is considerably larger than analyses in existing literature.
2. We highlight key challenges concerning cross-lingual citations that can inform future developments.
3. To facilitate further analyses, we make our collected data, the code used for analysis, and full results publicly available.[2]

The remainder of the paper is structured as follows. After briefly addressing our use of terminology down below, we give an overview of related work in Section 2.

[1] The selection of RQs is motivated by existing literature [18, 21] (1–3) as well as the intent to inform future endeavors in handling multilingual scholarly data (4–5).
[2] See https://github.com/IllDepence/icadl2020.

In Section 3 we discuss the identification of cross-lingual citations, data sources considered, and our data collection process. Subsequent analyses with regard to our research questions are then covered in Section 4. We end with a brief general discussion of our findings in Section 5, followed by concluding remarks in Section 6.

Terminology

Because *citation*, *reference* and related terms are not used consistently in literature, we shortly address their use in this paper. As shown in Figure 1, a cit*ing* document creates a bibliographical link to a cit*ed* document. We use the terms *citation* and *reference* interchangeably for this type of link (e.g., "*(vi)* in Figure 1 marks a cross-lingual reference," or "Paper^a makes two citations"). The textual manifestation of a bibliographic reference, often found at the end of a paper (e.g., "[1] Smith" in Figure 1), is referred to as *reference section entry*, or sometimes *reference* for short. We call the combined set of these entries *reference section*. Lastly, parts within the text of a paper, which contain a marker connected to one of the reference section entries, are called *in-text citations*.

2 Related Work

2.1 Cross-Lingual Citations in Academic Publications

Literature concerning cross-lingual citations in academic publications can be found in the form of analyses and applications. In [18] Kellsey and Knievel conduct an analysis of 468 articles containing 16,138 citations. The analysis spans 4 English language journals in the humanities (disciplines: history, classics, linguistics, and philosophy) over 5 particular years (1962, 1972, 1982, 1992, and 2002). The authors find that 21.3% of the citations in their corpus are cross-lingual, but note strong differences between the covered disciplines. Over time, they observe a steady total, but declining relative number of cross-lingual citations per article. The authors furthermore find, that the ratio of publications that contain at least one cross-lingual citation is increasing.

Lillis et al. [21] investigate if the global status of English is impacting the "citability" of non-English works in English publications. They base their analysison 240 articles from 2000 to 2007 in psychology journals, and furthermore use the Social Sciences Citation Index and ethnographic records. Their corpus contains 10,688 references, of which 8.5% are cross-lingual. Analyzing the prevalence of references in various contexts, they find that authors are more likely to cite a "local language" in English-medium national journals than in international journals. Further conducting analyses of e.g. in-text citation surface forms, they come to the conclusion that there are strong indicators for a pressure to cite English rather than non-English publications.

Similar observations are made by Kirchik et al. [20] concerning citations to Russian. Analyzing 498,221 papers in Thomson Reuters' Web of Science between

1993 and 2010, they find that Russian scholars are more than twice as likely to cite Russian publications when publishing in Russian language journals (21% of citations) than when they publish in English (10% of citations).

In [29] Schrader analyzes citations from non-English documents to English articles in open access and "traditional" journals. The corpus used comprises 403 cited articles published between 2011 and 2012 in the discipline of library and information science. The articles were cited 5,183 times (13.8% by non-English documents). In their analysis the author observes that being open access makes no statistically significant difference for the ratio of incoming cross-lingual citations of an article, or the language composition of citations a journal receives.

Apart from analyses, there are also approaches to prediction tasks based on cross-lingual citations [31, 15, 16, 25]. Tang et al. [31] propose a bilingual context-citation embedding algorithm for the task of predicting suitable citations to English publications in Chinese sentences. To train and evaluate their approach, they use 2,061 articles from 2002 to 2012 in the Chinese Journal of Computers, which contain citations to 17,693 English publications. Comparing to several baseline methods, they observe the best performance for their novel system. Similarly, in [15] and [16] Jiang et al. propose two novel document embedding methods jointly learned on publication content and citation relations. The corpus used in both cases consists of 14,631 Chinese computer science papers from the Wanfang digital library. The papers contain 11,252 references to Chinese publications and 27,101 references to English publications. For the task of predicting a list of suitable English language references for a Chinese query document, both approaches are reported to outperform a range of baseline methods.

In Table 1 we show a comparison of corpora between related work and our analysis.

Table 1 Comparison of corpora

Work	Type[a]	#Docs[b]	#Refs[b]	#Years	#Disciplines
Kellsey and Knievel [18]	en→*	468	16k	5[c]	4
Lillis et al. [21]	en→*	240	10k	7	1
Schrader [29]	*→en	403	5k	2	1
Tang et al. [31]	zh→en	2k	17k	10	1
Jiang et al. [15, 16]	zh→{en,zh}	14k	38k	n/a	1
Kirchik et al. [20]	{en,ru}→ru	497k	n/a	17	(unrestricted)
Ours	en→*	1.1M	39M	27	3

[a] type=focus reference type (en=English, ru=Russian, zh=Chinese, *=any)
[b] docs=documents, refs=references
[c] over a span of 40 years

2.2 Cross-Lingual Interconnections in Other Types of Media

Apart from academic publications, cross-lingual connections are also described in other types of media. Hale [11] analyzes cross-lingual hyperlinks between online blogs centered around a news event in 2010. In a corpus of 113,117 blog pages in English, Spanish, and Japanese, 12,527 hyperlinks (5.6% of them cross-lingual) are identified. Analysis finds that less than 2% of links in English blogs are cross-lingual, while the number in Spanish and Japanese blogs is slightly above 10%. Hyperlinks between Spanish and Japanese are almost inexistent (7 in total). Further investigating the development of links over time, the author observes a gradual decrease of language group insularity driven by individual translations of blog content—a phenomenon described as "bridgeblogging" by Zuckerman [34].

Similar structural features are reported by Eleta et al. [5] and Hale [10] for Twitter, where multilingual users are bridging language communities. As with academic publications, there also exists literature on link prediction tasks. In [17] Jin et al. analyze cross-lingual information cascades and develop a machine learning approach based on language and content features to predict the size and language distribution of such cascades.

3 Data Collection

3.1 Identification of Cross-Lingual Citations

Identifying cross-lingual citations requires information about the language of the citing and cited document, but this is often missing in scholarly data sets (cf. Table 2). Identifying the involved documents' language on the fly, however, is also challenging, because (a) full text (especially of cited documents) is not always available, and (b) language identification on short strings (e.g., titles in references) is unreliable [14]. To nevertheless be able to conduct an analysis of cross-lingual citations on a large scale, we utilize the practice of authors appending an explicit marker in the form of *"(in <Language>)"* to such references. This shifts the requirements from language metadata to the existence of (ideally unfiltered) reference section entries in the data.[3]

The question then remains, how common the practice of using such explicit markers is, compared to the use of untranslated non-English reference titles (without a marker). Conducting a comparison of both variants[4] on a random sample of one million reference section entries from the data set unarXive [28], we get a reliable estimate for non-Latin script languages (e.g., Chinese, Japanese, Russian), but inconclusive results for Latin script languages (e.g., German).[5]

[3] Language information is given for the cited document by the *"<Language>"* part of the marker, and for the citing document by the fact, that the marker is in English.

[4] Identification of marked entries is detailed in Section 3.3. For the identification of non-English titles we used the reference string parser module of GROBID [24] and the Python module langdetect (see https://github.com/Mimino666/langdetect).

[5] This is because the detection of untranslated non-English reference titles requires language identification on reference titles, which turned out to be unreliable for Latin script languages (e.g., many English titles were falsely identified as German).

Where we get reliable results, explicit marking appears to be the norm. In case of Russian, we observe 567 explicit markers and 3 untranslated titles without a marker. For Chinese, Japanese, and Greek, the number of explicit markers is 60, 57, and 7, respectively, compared to zero untranslated titles. Manual inspection of the noisy results for Latin script languages suggests a significant tendency toward using untranslated titles. These observations mean two things. First, a direct comparison between our numbers on non-Latin and Latin script languages is only valid for *explicitly marked* cross-lingual citations. Second, the number of undetected cross-lingual citations for non-Latin script languages such as Chinese, Japanese, and Russian, is negligible. Accordingly, concerning these languages, our results are valid for cross-lingual citations *in general*.

3.2 Data Source Selection

As our data source we considered five large scholarly data sets commonly used for citation related tasks [19, 7]. Table 2 gives an overview of their key properties. The Microsoft Academic Graph (MAG) and CORE are both very large data sets with some form of language metadata present. In the MAG the language is given not for documents themselves, but for URLs associated with papers. CORE contains a language label for 1.79% of its documents. S2ORC, the PubMed Central Open Access Subset (PMC OAS), and unarXive do not offer language metadata, but all contain some form of reference sections (GROBID [24] parse output, JATS [13] XML, and raw strings extracted from LaTeX source files respectively).

From these five, we decided to use unarXive and the MAG. This decision was motivated by two key reasons: (1) metadata of cited documents, and (2) evaluation of the "citability" of non-English works in English papers. As for (1), both S2ORC and the PMC OAS link references in their papers to document IDs within the data set itself (only partly in the PMC OAS, where also MEDLINE IDs and DOIs are found [9]). This is problematic in our case, because S2ORC is restricted to English papers, and the PMC OAS is constrained to Latin script contents,[6] which means metadata on non-English cited documents is inexistent (S2ORC) or very limited (PMC OAS). In unarXive, on the other hand, references are linked to the MAG, which contains metadata on publications regardless of language. Concerning reason (2), the fact that unarXive is built from papers on the preprint server arxiv.org, and the MAG contains metadata on paper's preprint *and* published versions, allows us to analyze whether or not cross-lingual citations are affected by the peer review process.

With these two data sources selected, the extent of our analysis is one million documents, across 3 disciplines (physics, mathematics, computer science), over a span of 27 years (1992–2019).

3.3 Data Collection

To identify references with *"(in <Language>)"* markers, we iterate through the total of 39.7M reference section entries in unarXive and first filter for the reg-

[6] See https://www.ncbi.nlm.nih.gov/pmc/about/faq/#q16.

Table 2 Overview of data sets

Data set	#D.[a]	Lang.meta.[b]	R.r.t.[c]	Reference sections	Used
MAG[d] [30, 33]	230M	(48%[e])	MAG	-	✓
CORE[f]	123M	1.79%	CORE	-	
S2ORC [23]	81M	-	S2ORC	34% (in GROBID parse)	
PubMed Central OAS[g]	2M	-	mixed	100% (in JATS XML)	
unarXive [28]	1M	-	MAG	100% (dedicated entity)	✓

[a] Number of documents
[b] Language metadata
[c] References resolved to
[d] Using version 2019-12-26
[e] Language given for source URLs (not always matching paper language)
[f] See https://core.ac.uk/. Using version 2018-03-01
[g] See https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

ular expression `\(\s*in\s+[a-zA-Z][a-z]+\s*\)`. This yields 51,380 matches with 207 unique tokens following *"in"* within the parentheses. Within these 207 tokens we manually identify non-languages (e.g., "press" or "preparation") and misspellings (e.g., "japanease" or "russain"), resulting in 44 unique language tokens. These are (presented in ISO 639-1 codes) be, bg, ca, cs, da, de, el, en, eo, es, et, fa, fi, fr, he, hi, hr, hu, hy, id, is, it, ja, ka, ko, la, lv, mk, mr, nl, no, pl, pt, ro, ru, sa, sk, sl, sr, sv, tr, uk, vi, and zh. These 44 languages cover 43 of the 78 languages, in which journals indexed in the Directory of Open Access Journals[7] (DOAJ) are published as of July 2020. The one language found in our data, but with no journal in the DOAJ, is Marathi. In terms of journal count by language, above 44 languages cover 97.54% of the DOAJ. In total, our data contains 33,290 reference section entries in 18,171 unique citing documents. We refer to this set of documents as the *cross-lingual set*.

To analyze differences between papers containing cross-lingual citations in unarXive and a comparable random set, we also generate a second set of papers. To ensure comparability we go through each year of the cross-lingual set, note the number of documents per discipline and then randomly sample the same number of documents from all of unarXive within this year and discipline. This means the *cross-lingual set* and the *random set* have the same document distribution across years and disciplines. Table 3 gives an overview of the resulting data used.

4 Results

In this section we describe the results of our analyses with regard to the research questions laid out in the introduction. We begin with general numbers indicating the prevalence of cross-lingual citations (based on unarXive alone) and follow with more in depth observations (utilizing the MAG metadata).

[7] See https://doaj.org/.

Table 3 Overview of data used

	Cross-lingual set	Random set	unarXive
#Docs	18,171	18,171	1,192,097
#Docs (MAG)	16,300	16,464	1,087,765
#Refs	635,154	536,672	39,694,083
#Refs (MAG)	290,421	242,090	15,954,664
#Cross-lingual refs	33,290	642	33,290

*docs=documents, refs=reference section entries, (MAG)=with a MAG ID.

4.1 Prevalence of Cross-Lingual Citations in English Papers

We find *"(in <Language>)"* markers in 33,290 out of 39,694,083 reference section entries (0.08%). These appear in 18,171 out of 1,192,097 documents (1.5%)—in other words in every 66th document. Of these 18k documents, 17,223 cite one language other than English, 864 cite two, 76 three, 7 documents four, and a single document cites works in English and five further languages (Russian, French, Polish, Italian, and German). The five most common language pairs within a single document are Russian-Ukrainian (277 documents), German-Russian (166), French-Russian (135), French-German (68), and Chinese-Russian (59). Table 4 shows the absolute number of reference section entries and unique citing documents for the five most prevalent languages, which combined make up over 90% in terms of both references and documents. As we can see, Russian is by far the most common, making up about two thirds of the cross-lingual set. When breaking down these numbers by year or discipline, it is important to also factor in the distribution of documents along these dimensions in the whole data set. Doing so, we show in Figure 2 the relative number of documents with cross-lingual citations over time for each of the aforementioned five languages. While the numbers in earlier years can be a bit unstable due to low numbers of total documents, we can observe a downwards trend of citations to Russian, an upwards trend of citations to Chinese, and a somewhat stable proportion in documents citing Japanese works. Looking at the numbers per discipline in Figure 3, we can see that cross-lingual citations occur most often in mathematics papers, and are about half as common in physics and computer science.

Table 4 Most prevalent languages

Language	#References	#Documents
Russian	23,922	12,304
Chinese	2,351	1,582
Japanese	1,843	1,397
German	1,244	965
French	931	719

Lastly, within the reference section of a document that has at least one cross-lingual citation, the mean value of "cross-linguality" (i.e., what portion of the reference section is cross-lingual) is 0.083 with a standard deviation of 0.099.

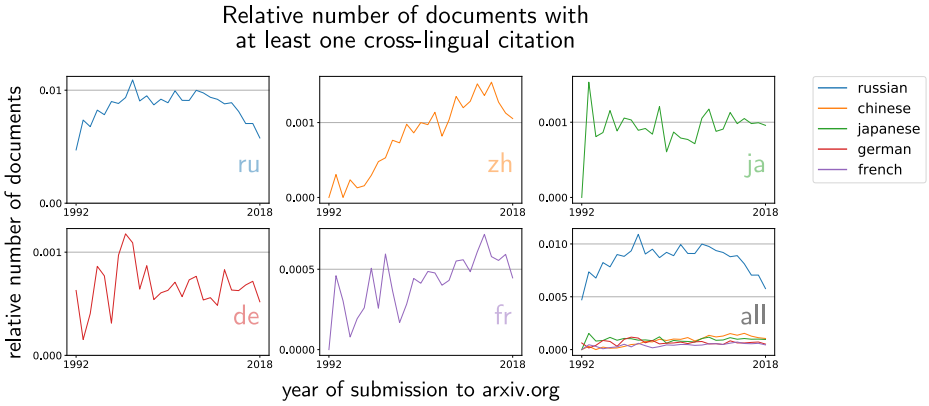

Figure 2 Relative number of documents citing Russian, Chinese, Japanese, German, and French works. Showing all aforementioned in the bottom right.

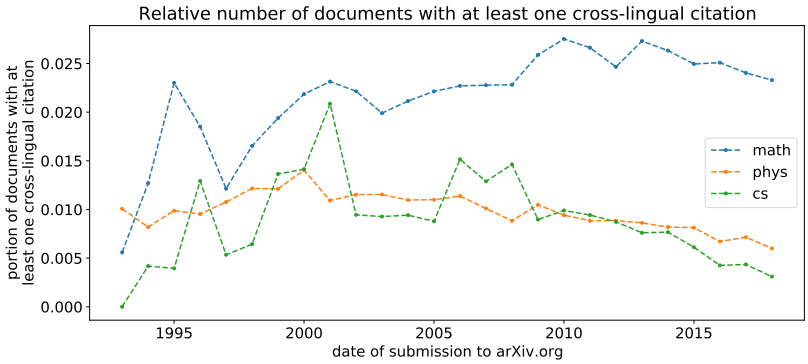

Figure 3 Relative number of mathematics, physics, and computer science documents citing non-English works.

Breaking these numbers down by discipline, we can see in Figure 4 that there is no large difference, although mathematics papers tend to have a slightly higher portion of cross-lingual citations. The mean values for mathematics, physics and computer science are 0.090, 0.078, and 0.080 respectively.

In terms of the prevalence of cross-lingual citations in English papers, we note that (in the disciplines of physics, mathematics and computer science) about 1 in 66 papers contains citations to non-English documents. About two thirds of these citations are to Russian documents, although in the last years there is a downwards trend with regard to Russian and an upwards trend in citations to Chinese. Citations to documents in Russian, Chinese, Japanese, German, and French make up 90% of the total of cross-lingual citations.

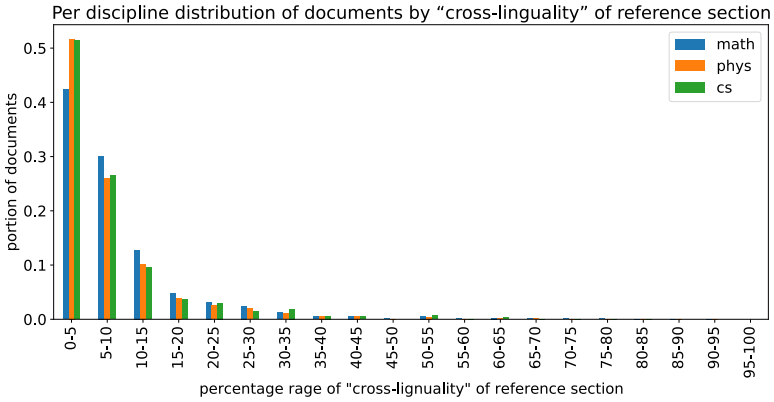

Figure 4 “Cross-linguality” of reference sections by discipline.

4.2 Impact of Cross-Lingual Citations in English Papers

As outlined in our research questions, apart from the prevalence of cross-lingual citations (RQ1), we also want to address whether or not self-citation is a driving factor (RQ2), if they are seen as an “acceptable” practice (RQ3), whether or not they pose a particular challenge for citation data mining (RQ4), and their potential impact on the success of the paper they’re part of (RQ5). Our results concerning these aspects are described in the following sections.

Self-citation To assess the relative degree of self-citation when referring to publications in other languages, we compare the ratio of self-citations in (a) the *cross-lingual citations* within the documents of the cross-lingual set, and (b) the *monolingual citations* within the documents of the cross-lingual set. Comparing two sets of citations from identical documents allows us to control for e.g. author specific self-citation bias. To determine self-citation, we rely on the author metadata in the MAG and therefore require both the citing and cited document of a reference to have a MAG ID. Within the cross-lingual set, this is the case for 3,370 cross-lingual references and 264,341 monolingual references. While at first, we strictly determined a self-citation by a match of MAG IDs, manual inspection of matches and non-matches revealed, that author disambiguation within the MAG is somewhat lacking—that is, in a non-trivial amount of cases there are several IDs for a single author. We therefore measure self-citation by two metrics. A strict metric which only counts a match of MAG IDs, and a loose metric which counts an overlap of the sets of author names on both ends of the reference as a self-citation.

Table 5 Self-citations

	Self-citations	
References to	loose	strict
non-English	19%	5%
English	17.9%	11.3%

Table 5 shows that going by the strict metric, self-citation is twice as common in monolingual citations. Applying the loose metric, however, self-citation appears to be slightly more common in cross-lingual citations. The larger discrepancy between the results of the strict and loose metric for cross-lingual citations suggests that authors publishing in multiple languages might be less well disambiguated in the MAG. With regard to self-citation being a motivating factor for cross-lingual citations—be it, for example, due to the need to reference one's own prior work—, we can note that this does not seem to be the case. Authors using cross-lingual citations appear to be at least equally as likely to self-cite when referencing English works.

"Acceptability" To assess the acceptance of cross-lingual citations by the scientific community—that is, whether or not non-English publications are deemed "citable" [21]—we analyze papers in our data that have both a preprint version as well as a published version (in a journal or conference proceedings) dated later than the preprint. This is the case for 2,982 papers. For each preprint-published paper pair, we check if there is a difference in cross-lingual citations. This gives an indication of how the process of peer review affects cross-lingual citations. We perform a manual as well as an automated analysis.[8]

For the manual evaluation, we take a random sample of 100 paper pairs. We then retrieve a PDF file of both the preprint and the published version, and manually compare their reference sections. For the automated evaluation, we find that 599 of the 2.9k paper pairs have PDF source URLs given in the MAG. After automatically downloading these and parsing them with GROBID, we are left with 498 valid sets of references. For these, we identify explicitly marked cross-lingual references as described in Section 3 and calculate their differences.

Table 6 shows the results of our evaluations. In both, cross-lingual citations are more often removed than added, but in the majority of cases left intact. The larger volatility in the automated evaluation is likely due to parsing inconsistencies of GROBID. Our findings complement those of Lillis et al. [21], who, analyzing psychology journals, observe *"some evidence that gatekeepers [...] are explicitly challenging citations in other languages."* For the fields of physics, mathematics, and computer science, we find no clear indication of a consistent in- or decreasing effect of the peer review process on cross-lingual citations.

Table 6 Changes in cross-ling. cit. between preprints and published papers

Evaluation	#Pairs	#Increased	#Deceased	Mean[a]	SD[a]
Manual	100	4	7	-0.02	0.529
Automated	498	33	70	-0.12	0.821

[a] of the differences in the amount of cross-lingual citations

[8] Full evaluation details can be found at https://github.com/IllDepence/icadl2020.

Impact on Citation Data Mining To assess if cross-lingual citations pose a particular challenge for scholarly data mining—and are therefore likely to be underrepresented in scholarly data—, we compare the ratio of references that could be resolved to MAG metadata records for the cross-lingual set and the whole unarXive data set. Of the 39M references in unarXive 42.6% are resolved to a MAG ID. For the complete reference sections of the papers in the cross-lingual set (i.e., references to both non-English and English documents) the number is 45.7% (290,421 of 635,154 references). Looking only at the cross-lingual citations, the success rate of reference resolution drops to 11.2% (3,734 of 33,290 references). We interpret this as a clear indication that resolving cross-lingual references is a challenge. Possible reasons for this are, for example:

1. A lack of language coverage in the target data set.
   For example, if the target data set only contains records of English papers, references to non-English publications cannot be found within and resolved to that target data set.
2. Missing metadata in the target data set.
   For example, when there is a primary non-English as well as an alternative English title of a publication, only the former is in the target data set's metadata, but the latter is used in the cross-lingual reference.
3. The use of a title translated "on the fly."
   If a non-English publication has no alternative English title, a self translated title in a reference cannot be found in any metadata. To give an example, reference [14] in `arXiv:1309.1264` titled *"Hierarchy of reversible logic elements with memory"* is only found in metadata[9(a)] as 記憶付き可逆論理素子の能力の階層構造について.
4. The use of a title transliterated "on the fly."
   Similar to an unofficial translated title, if a title is transliterated and this transliteration is not existent in metadata, the provided title is not resolvable. A concrete example of this is the third reference in `arXiv:cs/9912004` titled *"Daimeishi-ga Sasumono Sono Sashi-kata"* which is only found in metadata[9(b)] as 代名詞が指すもの, その指し方.

Cases 4 and especially 3 additionally impose a challenge on human readers, as the referred documents can only be found by trying to translate or transliterate back to the original. References to non-English documents which do not have an alternative English title should therefore ideally include enough information to (a) identify the referenced document (i.e., at least the original title), and (b) a way for readers not familiar with the cited document's language to get an idea of what is being cited (e.g., by adding a freely translated English title).[10] There are, however, situations where an original title cannot be used. Documents in PubMed Central, for example, cannot contain non-Latin scripts,[11] meaning that

[9] (a) http://hdl.handle.net/2433/172983 (b) https://ci.nii.ac.jp/naid/10008827159/.

[10] As, for example, in reference [15] in `arXiv:1503.05573`: "Шафаревич И. Р. Основы алгебраической геометрии// МЦНМО, Москва, 2007. (English translation: Shafarevich I.R. Foundations of Algebraic Geometry// MCCME, Moscow. 2007)."

[11] See https://www.ncbi.nlm.nih.gov/pmc/about/faq/#q16.

references to documents in Russian, Chinese, Japanese, etc. which do not have alternative English titles are inevitably a challenge for both human readers as well as data mining approaches, unless there is a DOI, URL, or similar identifier that can be referred to.

In light of this, taking a closer look at the 88.8% of unmatched references in the cross-lingual set broken down by languages, we note the following matching failure rates for the five most prevalent languages: Russian: 88.6%, Chinese: 87.0%, Japanese: 91.0%, German: 85.4%, and French: 83.2%. While all of these are high, the numbers for the three non-Latin script languages are noticeably higher than those of German and French. As can be seen with the task of resolving references—and as also indicated through our self-citation data shown in Table 5—cross-lingual citations do pose a particular challenge for scholarly data mining.

Impact on Paper Success To get an indication of whether or not an English paper's success is influenced by the fact that it contains citations to non-English documents, we compare our cross-lingual set with the random set (cf. Table 2). For both sets we first determine the number of papers that in the MAG metadata have a published version (journal or conference proceedings) in addition to the preprint on arxiv.org. That is, we assume that papers which only have a preprint version did not make it through the peer review process. Using this measure, we observe 9,390 of 16,224 (57.88%) successful papers in the cross-lingual set, and 10,966 of 16,378 (66.96%) successful papers in the random set. Unsurprisingly, due to the higher ratio of published versions, the papers in the random set are also cited more. Table 7 shows a comparison of the average number of citations that documents in both sets received. Due to the high standard deviation in the complete sets, we also look at papers which received between 1 and 100 citations, which are comparably frequent in both sets. As we can see, in the unfiltered as well as the filtered case, documents with cross-lingual citations tend to be cited a little less. Because here we can only control for the distribution of papers across years and disciplines, and not for individual authors (as we did in the "Self-citation" section), there might be various confounding factors involved.

Table 7 Comparison of citations received

Filter criterion		Cross-lingual set	Random set
-	#Docs	16,300	16,464
	Mean #cit	13.7	18.2
	SD	75.0	51.7
$1 \leq \#cit \leq 100$	#Docs	12,074	12,852
	Mean #cit	12.0	15.1
	SD	15.8	18.4

5 Discussion

Even though citations in English publications are typically to other English documents, we have seen that in preprints as well as conference proceedings and journal articles cross-lingual citations are used to refer to documents in a wide range of languages. Their prevalence is probably not high enough to greatly impact performance scores of general citation data driven approaches in e.g. information retrieval and recommendation—i.e., the evaluation of a system would not drastically change by introducing capabilities to handle references to other languages. However, as we could observe clear differences in prevalence across different disciplines and different cited languages, it might be advisable for specific approaches to evaluate the situation on a case by case basis. For example, a citation driven analysis of research trends in mathematics might benefit from being able to track "citation trails" into the realm of Russian publications.

We furthermore observed clear indicators that cross-lingual citations pose a challenge for citation data mining. As citation based performance evaluation is still a relevant steering mechanism in science, a lack in capabilities to automatically trace citations from e.g. international to national venues creates an imbalance between "supported" and "unsupported" publication languages. Furthermore, because some countries have sophisticated national systems and resources with regard to citation data—like Japan's CiNii[12] which has been used for research trend analysis [8]—, successful handling of cross-lingual citations would not just be a few additional data points on a subset of publications, but rather enable the detection of bridges between what are currently data silos that are not well interconnected.

6 Conclusion

Utilizing two large data sets, unarXive and the MAG, we performed a large-scale analysis of citations from English documents to non-English language works (cross-lingual citations). The data analyzed spans one million citing publications, 3 disciplines, and 27 years. We gain insights into cross-lingual citations' prevalence and impact, which we hope can inform further developments tackling the challenges of handling scholarly data.

Regarding English to non-English citations, we want to expand our investigation to further disciplines in the future. As our present analysis is based on papers in mathematics, physics, and computer science, insights into the humanities would be of particular interest. As for cross-lingual citations in general, analyses of non-English to English citations are likely to be more challenging to perform on a large scale, but might also yield insights with a larger impact, as citing English language publications is rather common in other languages, and has already given rise to approaches like cross-lingual citation recommendation.

[12] See https://support.nii.ac.jp/cia/cinii_db.

References

1. Abu-Jbara, A., Ezra, J., and Radev, D.: Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 596–606. Association for Computational Linguistics, Atlanta, Georgia (2013)
2. Chen, C.: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology **57**(3), 359–377 (2006). DOI: 10.1002/asi.20317
3. Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D.: SPECTER: Document-level Representation Learning using Citation-informed Transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2270–2282. Association for Computational Linguistics, Online (2020)
4. Colavizza, G., and Romanello, M.: Citation Mining of Humanities Journals: The Progress to Date and the Challenges Ahead. Journal of European Periodical Studies **4**(1), 36–53 (2019)
5. Eleta, I., and Golbeck, J.: Bridging languages in social networks: How multilingual users of Twitter connect language communities? Proceedings of the American Society for Information Science and Technology **49**(1), 1–4 (2012). DOI: 10.1002/meet.14504901327
6. Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., and Radev, D.: Blind men and elephants: What do citation summaries tell us about a research article? Journal of the American Society for Information Science and Technology **59**(1), 51–62 (2008)
7. Färber, M., and Jatowt, A.: Citation Recommendation: Approaches and Datasets. International Journal on Digital Libraries. (to appear)
8. Fukuda, S., Nanba, H., Takezawa, T., Takeda, H., Aizawa, A., Oumukai, I., Miyao, Y., and Uchiyama, K.: CiNii データベースを用いた研究動向分析システムの構築 (Construction of a CiNii database driven research trend analysis system). In: 言語処理学会第 18 回年次大会発表論文, pp. 539–542 (2012). (in Japanese)
9. Gipp, B., Meuschke, N., and Lipinski, M.: CITREC : An Evaluation Framework for Citation-Based Similarity Measures based on TREC Genomics and PubMed Central. In: iConference 2015 Proceedings. iSchools (2015)
10. Hale, S.A.: Global Connectivity and Multilinguals in the Twitter Network. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI ' 14, pp. 833–842. Association for Computing Machinery, Toronto, Ontario, Canada (2014). DOI: 10.1145/2556288.2557203
11. Hale, S.A.: Net Increase? Cross-Lingual Linking in the Blogosphere. Journal of Computer-Mediated Communication **17**(2), 135–151 (2012). DOI: 10.1111/j.1083-6101.2011.01568.x
12. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proceedings of the National academy of Sciences **102**(46), 16569–16572 (2005)
13. Huh, S.: Journal Article Tag Suite 1.0: National Information Standards Organization standard of journal extensible markup language. Science Editing **1**(2), 99–104 (2014). DOI: 10.6087/kcse.2014.1.99
14. Jauhiainen, T.S., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K.: Automatic language identification in texts: A survey. Journal of Artificial Intelligence Research **65**, 675–782 (2019)
15. Jiang, Z., Lu, Y., and Liu, X.: Cross-Language Citation Recommendation via Publication Content and Citation Representation Fusion. In: Proceedings of the

18th ACM/IEEE on Joint Conference on Digital Libraries. JCDL ' 18, pp. 347–348. Association for Computing Machinery, Fort Worth, Texas, USA (2018). DOI: 10.1145/3197026.3203898
16. Jiang, Z., Yin, Y., Gao, L., Lu, Y., and Liu, X.: Cross-Language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR ' 18, pp. 635–644. Association for Computing Machinery, New York, NY, USA (2018). DOI: 10.1145/3209978.3210032
17. Jin, H., Toyoda, M., and Yoshinaga, N.: Can Cross-Lingual Information Cascades Be Predicted on Twitter? In: Ciampaglia, G.L., Mashhadi, A., and Yasseri, T. (eds.) Social Informatics, pp. 457–472. Springer International Publishing, Cham (2017)
18. Kellsey, C., and Knievel, J.E.: Global English in the humanities? A longitudinal citation study of foreign-language use by humanities scholars. College & Research Libraries **65**(3), 194–204 (2004)
19. Khan, S., Liu, X., Shakil, K.A., and Alam, M.: A survey on scholarly data: From big data perspective. Information Processing & Management **53**(4), 923–944 (2017). DOI: 10.1016/j.ipm.2017.03.006
20. Kirchik, O., Gingras, Y., and Larivière, V.: Changes in publication languages and citation practices and their effect on the scientific impact of Russian science (1993–2010). Journal of the American Society for Information Science and Technology **63**(7), 1411–1419 (2012). DOI: 10.1002/asi.22642
21. Lillis, T., Hewings, A., Vladimirou, D., and Curry, M.J.: The geolinguistics of English as an academic lingua franca: citation practices across English-medium national and English-medium international journals. International Journal of Applied Linguistics **20**(1), 111–135 (2010). DOI: 10.1111/j.1473-4192.2009.00233.x
22. Liu, X., and Chen, X.: CJK Languages or English: Languages Used by Academic Journals in China, Japan, and Korea. Journal of Scholarly Publishing **50**(3), 201–214 (2019)
23. Lo, K., Wang, L.L., Neumann, M., Kinney, R., and Weld, D.: S2ORC: The Semantic Scholar Open Research Corpus. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4969–4983. Association for Computational Linguistics (2020)
24. Lopez, P.: GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In: Research and Advanced Technology for Digital Libraries, pp. 473–474 (2009)
25. Ma, S., Zhang, C., and Liu, X.: A review of citation recommendation: from textual content to enriched context. Scientometrics **122**(3), 1445–1472 (2020)
26. Moed, H.F., Markusova, V., and Akoev, M.: Trends in Russian research output indexed in Scopus and Web of Science. Scientometrics **116**(2), 1153–1180 (2018)
27. Moskaleva, O., and Akoev, M.: Non-English language publications in Citation Indexes - quantity and quality. In: Proceedings 17th International Conference on Scientometrics & Informetrics, pp. 35–46. Edizioni Efesto, Italy (2019)
28. Saier, T., and Färber, M.: unarXive: a large scholarly data set with publications' full-text, annotated in-text citations, and links to metadata. Scientometrics (2020). DOI: 10.1007/s11192-020-03382-z
29. Schrader, B.: *Cross-language Citation Analysis of Traditional and Open Access Journals*, (2019). DOI: 10.17615/djpr-1k06. Feb. 2019
30. Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J.(, and Wang, K.: An Overview of Microsoft Academic Service (MAS) and Applications. In: Proceedings

of the 24th International Conference on World Wide Web. WWW '15 Companion, pp. 243–246. ACM (2015). DOI: 10.1145/2740908.2742839
31. Tang, X., Wan, X., and Zhang, X.: Cross-Language Context-Aware Citation Recommendation in Scientific Articles. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval. SIGIR '14, pp. 817–826. Association for Computing Machinery, New York, NY, USA (2014). DOI: 10.1145/2600428.2609564
32. Vera-Baceta, M.-A., Thelwall, M., and Kousha, K.: Web of Science and Scopus language coverage. Scientometrics **121**(3), 1803–1813 (2019)
33. Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A., and Rogahn, R.: A Review of Microsoft Academic Services for Science of Science Studies. Frontiers in Big Data **2**, 45 (2019). DOI: 10.3389/fdata.2019.00045
34. Zuckerman, E.: Meet the bridgebloggers. Public Choice **134**(1), 47–65 (2008)