Citing Foreign Language Sources : an Analysis of the S2ORC Dataset

Marc Bertin, Iana Atanassova

Citing Foreign Language Sources : an Analysis of the S2ORC Dataset

Marc Bertin[1,*,†], Iana Atanassova[2,†]

[1]*Université Claude Bernard Lyon 1, ELICO*
43 Boulevard du 11 novembre 1918 69622 Villeurbanne cedex, France
[2]*Université de Franche-Comté, CRIT*
30 rue Mégevand, F-25000 Besançon, France
Institut Universitaire de France (IUF), France

Résumé

In this article we investigate the multilingualism of references in the Semantic Scholar Open Research Corpus (S2ORC). While this dataset contains peer-reviewed papers from different disciplines, written mainly in English, we identify the languages used in the references, their linguistic groups and their distribution over time. The results allow us to observe the dominance of English in science, as well as the relative proportions of the other 34 languages, representing over 2.4 million of the cited sources, and their linguistic groups. We show that the relative share of non-English citations has been increasing since 2000. However, the vast majority of citations in non-English publications are to English sources. We discuss some of the limitations of this study, mainly related to the nature of the dataset, which is biased towards English, and the quality of the language detection tool.

Keywords

Foreign language sources, Language detection, English, S2ORC, Bibliographic References, Bibliometrics

1. Introduction

The English language has gradually gained a dominant position in science [1]. At the same time, the question of the use of different languages in publications remains a subject of scientific debate, which also finds an echo at the political level. It has been shown by Kirillova in 2019 [2] that the publication of articles in English in a given country is related to its scientific activity as well as to the size of the country. These results reflect the desire of large countries to maintain their national language as the language of science, while small non-English speaking countries aim to reach international standards.

. *BIR 2023 : 13th International Workshop on Bibliometric-enhanced Information Retrieval at ECIR 2023, April 2, 2023*

*. Corresponding author.
†. These authors contributed equally.
. marc.bertin@univ-lyon1.fr (M. Bertin) ; iana.atanassova@univ-fcomte.fr (I. Atanassova)
. https://elico-recherche.msh-lse.fr/membres/marc-bertin (M. Bertin) ; iana-atanassova.github.io/ (I. Atanassova)
. 0000-0003-1803-6952 (M. Bertin) ; 0000-0003-3571-4006 (I. Atanassova)

In addition, Moskaleva and Akoev [3] showed that non-English native language publications are less read and cited than those in English outside the home country. They also observed that the ranking of journals is correlated with the share of English publications for multilingual journals. In terms of content, Di Bitetti compared the abstracts of a bilingual journal showing that there were no differences in aspects such as quality, general interest, etc. between articles published in English and in other languages in these journals [4].

The work of Smirnova and Lillis in 2022 is based on a corpus comparing research articles written in Russian with those written in English in the following disciplines : philosophy, sociology and economics [5]. At the micro level, the article analyses the changes in citations in the English and Russian texts. At the macro level, the article raises questions about what is considered ”citation-worthy” in different geolinguistic contexts and considers the consequences of citation brokerage and knowledge production practices and circulation on a global scale. The work of Angulo en 2021 [6] shows that science based on a single language, particularly English, can be a barrier to knowledge transfer that can lead to bias in the provision of global models. Experience shows that including non-English sources can reduce bias in understanding and enrich scientific knowledge. Faced with this hegemony, some publishers have begun to advocate bilingual journals [7].

In this article we address the problem of multilingualism in publications through a corpus-based experiment using the Semantic Scholar Open Research Corpus (S2ORC) [8]. Our objective is to examine foreign language (non-English) references in this corpus, and observe their nature and distribution.

S2ORC is a large dataset of full-text peer-reviewed research articles, mainly written in English. We chose this dataset for our experiment because it focuses on English research articles and spans many academic disciplines. S2ORC includes articles from a wide range of scientific fields including medicine, biology, chemistry, engineering, computer science, physics, materials science, mathematics, psychology, economics, political science, business, geology, sociology, geography, environmental science, art, history and philosophy. Other large datasets of research articles exist, such as the thematic dataset COVID19 [9], Climate change [10] or the multilingual dataset ISTEX [11], which have different coverage.

When studying references to foreign language sources, we should take into account the problem of the multiplicity of scripts. Indeed, if we look at an article written in English, it may contain references in other languages that are expressed either in the native alphabet of the foreign language or in Latin characters. Traditionally, two types of operations have been used to romanise languages with non-Latin scripts. On the one hand, transliteration is the operation that consists in replacing each grapheme of one writing system by a grapheme or group of graphemes of another system, regardless of the pronunciation. On the other hand, transcription is the opposite of transliteration. In transcription, each phoneme of a language is replaced by a grapheme or group of graphemes from one writing system to another. Transliteration is based on national and international standards [1].

1. E.g. Armenian : ISO 9985 : 1996 ; Macedonian, Turkish, Russian, Ukrainian, Belarusian, Bulgarian, non-Slavic languages in Cyrillic, based on ISO 9 : 1995 ; Chinese with NF ISO 7098 : 1992 ; Georgian, using ISO 9984 : 1996 ; Greek with ISO 843 : 1997 ; Hebrew, Yiddish or Syriac, which is a Hebrew script based on the NF ISO 259-2 : 1995 standard ; or Thai, which uses the ISO 11940 : 1998 standard.

2. Methods

2.1. Dataset

We use the full S2ORC version 1 dataset, which contains approximately 81 million open access articles published up to 2020. While the dataset is intended to include only English language articles, a closer look reveals that a small proportion of the articles are in other languages. We show this below. The dataset is available in json format. Each article is identified by its `paper_id`. For our experiment, we extracted the metadata of articles (title and year) and the metadata of the bibliographic references they contain (titles, years, journals, etc.).

2.2. Processing Pipeline

The language of each article and bibliographic reference was detected using Google's Compact Language Detector v3 (`gcld3`) Python library [2], which implements a neural network model for language identification. We used the title of the reference as input. Several outputs are provided by `gcld3`, including the code [3] of the most likely language and the likelihood score for that language. The titles of the bibliographic references were used for the language detection.

In order to obtain a good quality sample, we discarded references for which the probability score was lower than 0.95. This is the case, for example, when the title is too short to identify the language with certainty. References with missing metadata, e.g. missing year, were also ignored (less than 1 % of all references). Furthermore, papers and references with a year before 1950 or after 2020 were ignored, as such values are most likely due to typing errors in the dataset (less than 0.5% of all references).

The metadata for the citing paper (title and year) was obtained by using its `paper_id` in S2ORC. The language of the citing paper was determined in the same way as for the references, using its title.

The quality of the language detection varies for some of the poorly endowed languages. The choice of the `gcld3` library was done after testing several other libraries for language detection, e.g. `spacy lang-detect`. We found that for our task, `gcld3` provides better results and is much faster. We manually evaluated language detection by examining a subset of 100 titles for each language. If the observed precision was below 50 % we excluded that language [4] from our experiment. For the majority of the languages that remained the precision is above 75%.

The `gcld3` recognises cases of Latin transliteration and assigns a language code followed by `-Latn`, e.g. `ru-Latn` stands for Russian text transliterated with Latin characters. In our experiment, we do not make a distinction between references that are transliterated and those that are written in the native script.

2. https://github.com/google/cld3

3. The language codes use the IETF BCP 47 language tags, which combine several standards : ISO 639, ISO 15924, ISO 3166-1 and UN M.49. F. The subtags are maintained by the IANA Language Subtag Registry.

4. This applies to the following languages : `af, ar, az, be, bg-Latn, ceb, co, cy, eo, et, eu, fil, fy, ga, gd, gl, ha, haw, hi, ht, ig, jv, kk, ku, ky, lb, mg, mi, mk, mn, ms, mt, ne, ny, sm, sn, so, sq, st, su, sw, uz, vi, xh, yi, yo, zu`.

Table 1
Linguistic Groups and Language Tags

Linguistic Group	IETF BCP 47 language tag
Afro-Asiatic Cushitic	so
Afro-Asiatic Semitic	ha,mt
Austroasiatic	ceb,fil,haw,id,jv,mg,mi,ms,sm,su,tl,vi
Auxiliary (Esperanto)	eo
Basque	eu
Central Semitic	ar,he,iw
Creole	ht
English	en
Indo-European Albanian	sq
Indo-European Baltic	lt,lv,
Indo-European Celtic	cy,ga,gd
Indo-European Germanic	af,da,de,fy,is,lb,nl,no,sv,yi
Indo-European Hellenic	el,el-Latn,gr
Indo-European Indo-Iranian	bn,hi,ku,ne,tg
Indo-European Romance	ca,co,es,fr,gl,it,la,pt,ro
Indo-European Slavic	be,bg,bg-Latn,cs,cz,hr,mk,pl,ru,ru-Latn,sk,sl,sr,uk
Japanese	ja,ja-Latn
Koreanic	ko
Mongolian	mn
Niger-Congo	ig,ny,sn,st,sw,xh,yo,zu
Sino-Tibetan Sinitic	zh,zh-Latn
Thai	th
Turkic	az,kk,ky,tr,uz
Uralic Finno-Ugric	et,fi,hu
West Semitic	am

2.3. Composition of the Linguistic Groups

In order to study how the different language groups are represented in the dataset, we have classified the languages with their language tags into language groups. The summary is presented in table 1, which contains all language tags identified as present in the dataset. Only English was not associated with its group (the Indo-European Germanic languages), but we have separated it from the other languages because it is the dominant language of the corpus. Languages that were excluded from the experiment because of the poor quality of the language detection are presented in red.

3. Results

As a result of the language detection, after applying the above criteria, we obtained a subset of S2ORC containing a total of 5.9 million articles with 109.9 million references for which the language was detected with a probability score above 0.95.

3.1. Languages present in the corpus

Although the dataset is intended to contain only English research articles, we found 35,463 articles that were identified as being written in other languages, out of a total of 5,934,799 articles (0.60 %). Figure 1 shows a bar chart of the number of articles found for each language.

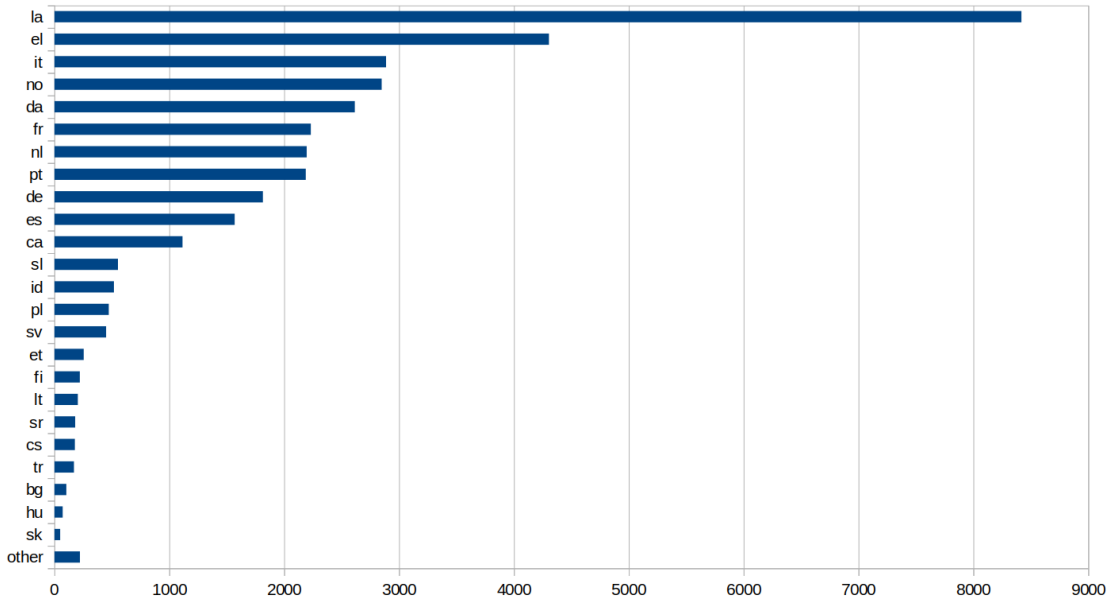

Figure 1 : Non-English articles in the S2ORC dataset

Most of the articles are in Latin, and this result may be biased because titles in the biomedical domain may be misclassified as Latin text. For the remaining languages, we manually sampled and checked the presence of such articles in the dataset. In our study, we have taken this result as an opportunity to work with a subset of multilingual (non-English) research articles and study their references. Overall, we found that the dataset contains articles in 32 languages (including English) and references to sources in 35 languages (including English).

As might be expected, the vast majority of references in the dataset are to English sources. They account for 107,437,865 references out of a total of 109,838,938 references (97.81 %). The remaining 2,401,073 references are to non-English sources. Figure 2 shows a bar chart of the number of references to sources in the different languages.

To gain a better understanding of the types of foreign language sources cited, we extracted titles of non-English sources cited in English articles and translated some of them for analysis and discussion. The translation is intended to be informative. The results are shown in the table 2. Languages that were excluded from the experiment because of the poor quality of the language detection are presented in red. We can see that the titles correspond to studies that have local significance and dimension, relate to a particular country or geographical area, and are written in their native language. For example, the title in haw (Hawaiian) represents a study on the history of "*Kahalu'u* and *Keauhou*". The example in sq (Albanian) is a study of

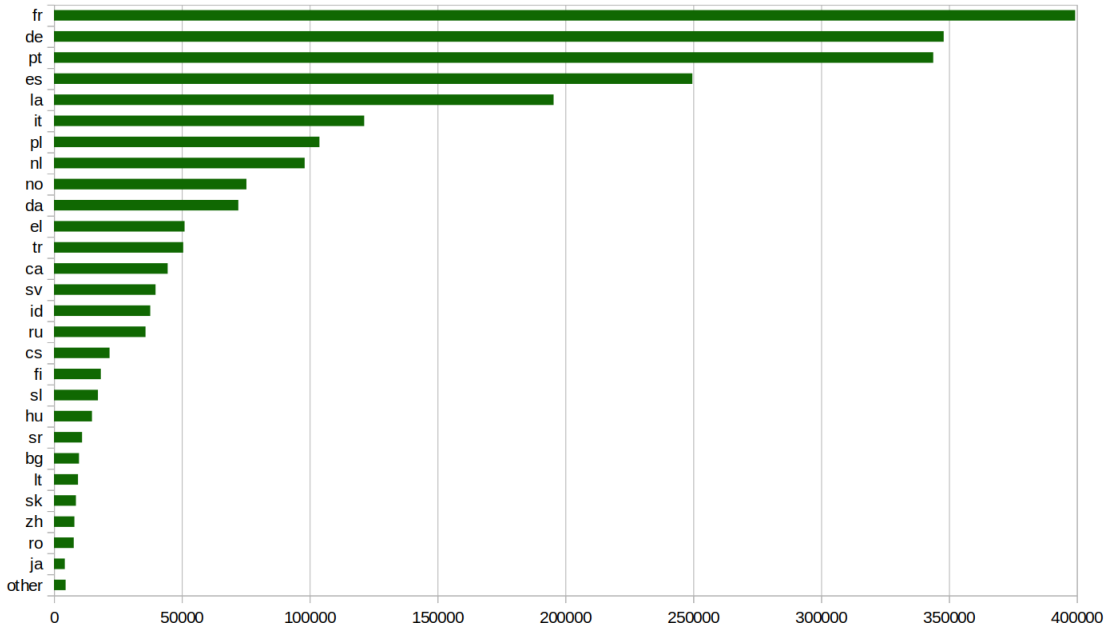

Figure 2 : References to non-English sources in the S2ORC dataset

hydrocarbon energy in Albania, and the example in `mk` (Macedonian) is about geological studies in the villages of *Kosel* and *Pesočan*.

3.2. Evolution over time

We have studied the distribution of citations to non-English sources with respect to the year of reference and with respect to the year of publication of the citing article.

Figure 3 shows the relative proportion of non-English sources cited by articles published between 1950 and 2020. The numbers on the horizontal axis represent the average number of sources cited per article in our dataset. The nine most frequent languages are shown in different colours, and the last group contains all other languages. We can see that between 1960 and 1980, the non-English references cited were mostly in French and German. Since 2000, however, their share has been decreasing, with the appearance of a large number of sources cited in Portuguese and Spanish. Overall, the relative share of non-English references has increased since 2000.

3.3. Use of English in relation with other languages

To study how foreign language sources are cited in English articles, we looked at the subset of references from English articles to non-English sources. Figure 4 shows the relative proportions of different language groups cited in English articles. The left-hand side of the figure shows the linguistic groups of the citing articles, while the right-hand side shows the linguistic groups of the references. The majority of references are from the Indo-European Romance and Germanic

Table 2
Examples of titles for non-English references

Lang. tag	Title of article
haw	He Mo'olelo 'Äina : Kahalu'uKaulana i ka Wai Puka iki o Helani a me Keauhou-I ka 'Ili'ili Nehe : A History of Kahalu'u and Keauhou Ahupua'a District of Kona, Island of Hawai'i. Kumu Pono Associates LLC
transl.	A History of Kahalu'u and Keauhou Ahupua'a District of Kona, Island of Hawai'i. Teacher Right Associates LLC
hi-Latn	Dolpa Jillama Yarsagumba Sankalan Tatha Byebasthapan Ek Parichaye
transl.	Yarsagumba collection and settlement an introduction in Dolpa district
is	Sjávarnytjar við Ísland. Mál og menning
transl.	Seafood off Iceland. Language and culture
lt	Lietuvos nekilnojamojo turto rinka : nekilnojamojo turto ir sta tybos sąnaudų kainų analizė
transl.	Lithuanian real estate market : analysis of real estate and construction cost prices
lv	Latviešu tēlotāja māksla 1860-1940. Rīga : Zinātne
transl.	Latvian painter's art 1860-1940. Riga : Science
mi	Te Poha o Tohu Raumati Te Rūnanga o Kaikōura Environmental Management Plan. Te Rūnanga o Kaikōura, Takahanga Marae Kaikōura
transl.	Te Poha o Tohu Rāmati Te Rūnanga of Kaikōura Environmental Management Plan. The Kaikōura Cabinet, Kaikōura Field Events
mk	Геолошки извештај за сулфатните појави на селата Косел, Песочани. Вапила и Влгоште. Стручен фонд на Геолошки завод -Скопје
transl.	Geological report on sulfate occurrences in the villages of Kosel, Pesočani. Vapila and Vlgoshte. Professional fund of Geological Institute - Skopje
ro	Zece mii de culturi, o singură civilizaţie. Spre geomodernitatea secolului XXI
transl.	Ten thousand cultures, one civilization. Towards the geomodernity of the 21st century
ru-Latn	Nauchno-tekhnologicheskie prioritety dlya modernizatsii rossiyskoy ekonomiki S&T Priorities for Modernization of Russian Economy
transl.	Scientific and technological priorities for the modernization of the Russian economy S&T Prioritize for Modernization of Russian Economy
sk	Vecná a časová zmena motivácie riadiacich zamestnancov v Slovenských elektrárňach a. s. Mochovce. On-line odborný časopis Manažment v teórii a praxi
transl.	Material and temporal change in the motivation of management employees in Slovenské elektrárňa a. with. Mochovce. On-line professional magazine Management in theory and practice
sq	Konsumi i energjisë së hidrokarbureve në Shqipëri dhe në Botë në vitet
transl.	Energy consumption of hydrocarbons in Albania and in the world in years
sw	Anayedhulimiwa asipopambana na huyo dhalimu, yeye ataendelea kuteseka wakati dhalimu atastarehe kwa amani', An-Nuur
transl.	If the oppressed does not fight the oppressor, he will continue to suffer while the oppressor rests in peace', An-Nuur
tg	Садриддин Айнӣ ва баъзе масъалаҳои инкишофи забони адабии тоҷик
transl.	Sadriddin Ainy and some issues of Tajik literary language development
uk	Напрями формування інноваційної системи нового технологічного укладу в Україні
transl.	Directions of the formation of the innovative system of the new technological order in Ukraine
uk	Стан корупції в Україні. Порівняльний аналіз загальнонаціональних досліджень
transl.	The state of corruption in Ukraine. Comparative analysis of national studies
vi	Những cây thuốc và vị thuốc Việt Nam, Nhà xuất bản Khoa học Kỹ thuật
transl.	Vietnamese medicinal plants and herbs, Science and Technology Publishing House
zh-Latn	Jiating yinsu dui weichengnianren fanzui de yingxiang ji duice shizhengyanjiu (Empirical study on the impacts of family factor on juvenile delinquency and solution : A case study in Chongqing)
transl.	An Empirical Study on the Influence of JIA Listening Factors on Juvenile Delinquency and Countermeasures

groups, which are the closest languages to English. The Indo-European Slavic languages come next, and all the other groups have very small proportions of citations.

Finally, we examined the subset of references from non-English articles. The results are shown in figure 5. Again, the left side of the figure represents the language groups of the citing articles, while the right side represents the language groups of the references. It is interesting to note the dominance of English in this group, as we see that the majority of citations in these articles

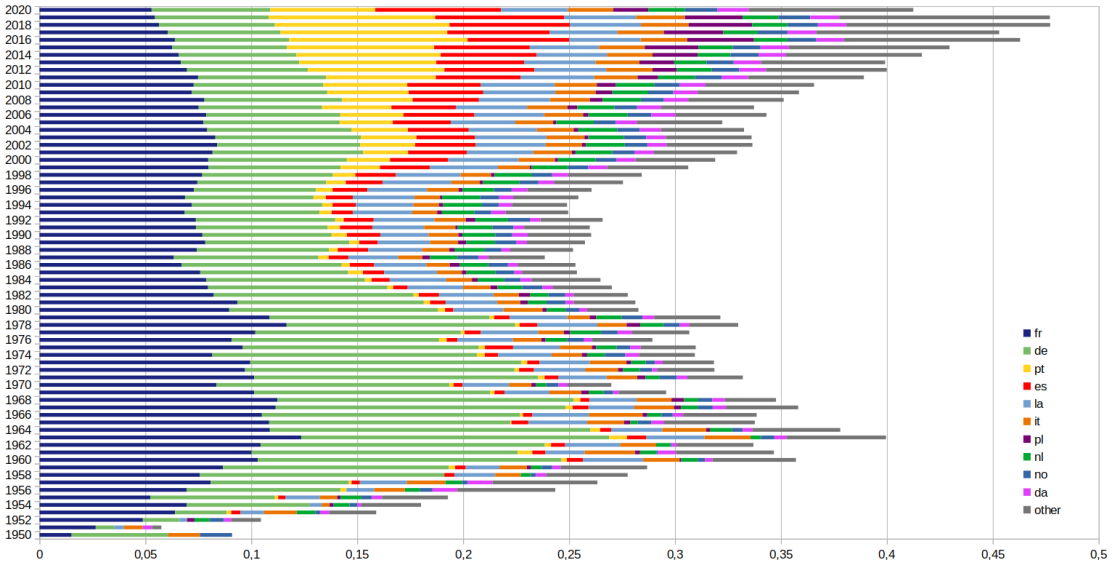

FIGURE 3 : Relative part of non-English sources with respect to the year of the publication

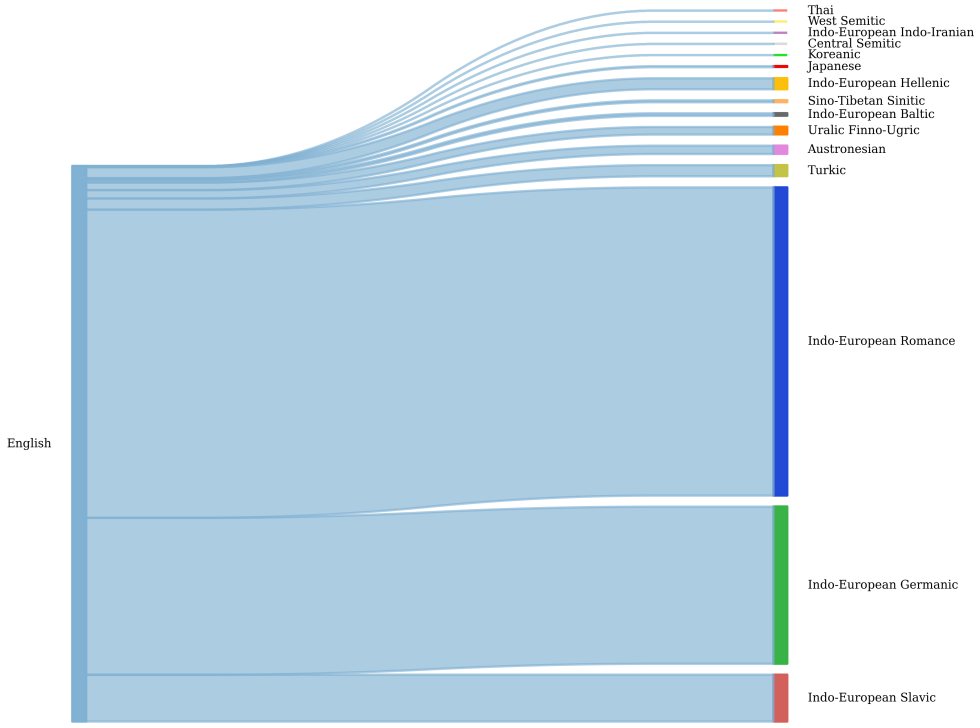

FIGURE 4 : English articles citing non English sources

are to English sources. In addition, articles written in all other language groups cite a majority of English sources. In particular, a significant proportion of publications in Indo-European Romance languages (other than English) cite sources in the same language group, and the same is true for the other Indo-European groups.

Figure 5 : References in non English articles

4. Discussion and Limitations

The results presented in this study may be biased by several different factors. Firstly, the quality of language detection may vary between different languages, particularly for poorly endowed languages which may have a low recognition rate. Manual cleaning and evaluation would be required to produce better quality data. Secondly, the use of titles as proxies to detect the language of a publication is not optimal, as titles can be very short or contain technical terms that may wrongly point to Latin or English. However, given the size of the dataset and the data available for references, this was the only way to identify the languages of the sources. Another possibility would be to use titles and abstracts. This study could therefore be improved by harvesting abstracts for the references in the dataset.

The choice of the S2ORC dataset introduces an important bias related to the English language. As S2ORC is intended to include only English language research articles, the subset of non-English language articles that we analysed may not be representative of these languages, as it is

made up of publications that are included in S2ORC. The S2ORC dataset was constructed by retaining only papers identified as English using the `cld2` tool run over titles and abstracts [8]. This introduces a bias towards English and other languages that use the Latin alphabet. We can suppose that such a tool is likely to perform better when it comes to excluding all non-latin non-English articles from the dataset than the articles written in Latin script. This may lead to an over-representation of Indo-European Romance languages. Furthermore, as the extraction pipeline used for S2ORC is optimised for the English language, and the extraction of citations to foreign language papers has not been evaluated for the creation of the dataset. Thus, it can be expected that the extracted references are biased towards English language references.

In general, the choice of sources to be cited in a publication and their languages is a complex process influenced by many factors, such as the languages spoken by co-authors, the subject of the study, and so on. The editorial requirements of some journals may favour English language references or require the titles of foreign language references to be translated into the language of the publication.

5. Conclusion and Perspectives

We carried out an analysis of the Semantic Scholar Open Research Corpus (S2ORC) and identified the languages of the research articles and their references based on their titles. While the vast majority are in English, we found articles in 44 different languages and references in 54 languages. We observed their linguistic groups and their distribution over time between 1950 and 2020. The results allow us to observe the dominance of English in science, where the vast majority of citations in non-English publications are to English sources. We also show that the relative share of non-English citations will increase from 2020 onwards.

Evaluating and improving the quality of language detection can provide us with more reliable results in the future. In the long run, this work aims to propose a typology of foreign language references in order to better understand the way they are used in publications. Indeed, one of the issues that interest bibliometricians today is the reflection on the context of citation. The classification of citation contexts according to multilingual criteria has not been addressed in the literature to our knowledge. To this end, an extension of this approach will focus on national and other multilingual corpora. It also seems relevant to study such references in terms of their place in publications and their linguistic contexts.

Acknowledgments

This work was supported by grant number ANR-20-CE38-0003-01 and grant number ANR-21-CE38-0003-01.

Références

[1] P. S. Rao, The role of english as a global language, Research Journal of English 4 (2019) 65–79.

[2] O. V. Kirillova, Publication language and the journal scientometric indicators in global citation databases, Science Editor and Publisher 4 (2019) 21–33. doi :10.24069/2542-0267-2019-1-2-21-33.

[3] O. Moskaleva, M. Akoev, Non-english language publications in citation indexes - quantity and quality, CoRR abs/1907.06499 (2019). URL : http://arxiv.org/abs/1907.06499.

[4] M. S. Di Bitetti, J. A. Ferreras, Publish (in english) or perish : The effect on citation rate of using languages other than english in scientific publications, Ambio 46 (2017) 121–127. doi :10.1007/s13280-016-0820-7.

[5] N. Smirnova, T. Lillis, Citation in global academic knowledge making : A paired text history methodology for studying citation practices in english and russian, Journal of English for Research Publication Purposes 3 (2022) 78–108.

[6] E. Angulo, C. Diagne, L. Ballesteros-Mejia, T. Adamjy, D. A. Ahmed, E. Akulov, A. K. Banerjee, C. Capinha, C. A. Dia, G. Dobigny, V. G. Duboscq-Carra, M. Golivets, P. J. Haubrock, G. Heringer, N. Kirichenko, M. Kourantidou, C. Liu, M. A. Nuñez, D. Renault, D. Roiz, A. Taheri, L. N. Verbrugge, Y. Watari, W. Xiong, F. Courchamp, Non-english languages enrich scientific knowledge : The example of economic costs of biological invasions, Science of The Total Environment 775 (2021) 144441. doi :10.1016/j.scitotenv.2020.144441.

[7] M. D. Rosselli, Moving towards english, Acta Neurológica Colombiana 36 (2020) 1–2. doi :10.22379/24224022270.

[8] K. Lo, L. L. Wang, M. Neumann, R. Kinney, D. Weld, S2ORC : The semantic scholar open research corpus, in : Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4969–4983. doi :10.18653/v1/2020.acl-main.447.

[9] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19 : The COVID-19 open research dataset, in : Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020. URL : https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.

[10] R. Grundmann, R. Krishnamurthy, The discourse of climate change : A corpus-based approach, Critical approaches to discourse analysis across disciplines 4 (2010) 125–146.

[11] P. Cuxac, A. Collignon, ISTEX, un projet national d'archives documentaires : au-delà de l'accès au texte intégral, l'enrichissement des données par méthodes de fouille de textes, in : Analyser la science : les bibliothèques numériques comme objet de recherche in 85ème Congrès ACFAS, Montréal, Canada, 2017. URL : https://hal.science/hal-01869036.