The patterns of publication languages used by STEM research in China

Kai Li[1], Jiajing Chen[2] and Chaoqun Ni[3]

[1] *kli16@utk.edu*
University of Tennessee at Knoxville, School of Information Sciences, 1345 Circle Park Drive 451 Communications Building, Knoxville, TN, 37996 (United States)

[2] *jchen995@wisc.edu*
New York University, Department of Computer Science, 251 Mercer Street, New York, NY, 10012 (United States)

[3] *chaoqun.ni@wisc.edu*
University of Wisconsin-Madison, Information School, 600 N Park St #4217, Madison, WI, 53706 (United States)

Abstract

English has become the international language of science, especially in China. Researchers in China have been increasingly publishing in the English language, as evidenced by the exponential growth of publications by China in major bibliographic databases in recent decades. Nevertheless, there has been limited research on the exact pattern of languages used by Chinese researchers due to a lack of infrastructure covering publications in both English and Chinese produced in China. In this study, we present preliminary findings from our project analyzing all research outputs funded by the National Science Foundation of China (NSFC), the leading government research funder in natural sciences and engineering in China. We strive to understand what languages Chinese STEM researchers use in journal and conference publications, how language use varies by research domains and projects, and how the trends evolve. Our preliminary results show that English was the most popular language in STEM research in the 2010s, and its popularity increased during this decade. Moreover, the use of the English language is also positively correlated with the seniority of project PI as embedded in project classes. Our preliminary results provide the basis for more granular analyses and more constructive conversations about the use of languages in scientific research in China.

Introduction

Using various languages in scientific publications is critical to constructing an inclusive, dynamic, and effective global research system. Despite the dominant power of the English language in the circulation of knowledge in the global science system, other languages are frequently used by researchers in countries where English is not the only language (González-Alcaide et al., 2012). In fact, multilingual publishing plays a central role in keeping research concerning local context and cultural heritage alive (Kulczycki et al., 2020).

Nevertheless, multilingualism in scientific publications has long been a challenge for scientometric scholarship and global research policies. It has not been easy to consider publications in multiple languages in analyzing researchers, fields, or other entities. Large-scale scientometric studies typically rely on bibliographic databases, such as Web of Science (WoS) and Scopus. Nevertheless, they typically suffer from the overrepresentation of English-language publications across knowledge domains (Mongeon & Paul-Hus, 2016; van Leeuwen et al., 2001). Given the vital relevance of local languages to native speakers and local journals (Liu et al., 2018), researchers call for a more balanced approach to multilingualism in scholarly communication (Sivertsen, 2018).

Multilingual science is also a global phenomenon. In some European countries where English is not the official language, many researchers publish in their local languages, especially those in social sciences and humanities (Kulczycki et al., 2018, 2020; Mathies et al., 2020; Sivertsen

& Larsen, 2012). As the new powerhouse of scientific publication, China publishes substantially in languages other than Chinese, mainly English, due to various incentives for publishing in English. Meanwhile, the role that publications in Chinese play in the national research system should not be overlooked (Liu et al., 2018). However, there has been limited research (Zhou et al., 2010) on the exact pattern of languages used by the research output of China. We argue that this issue is particularly relevant in light of China's new national research policies, which focus on using local languages to address local issues (Liang et al., 2022). This is a significant shift from the policies that strongly favor Science Citation Index (SCI) and English-language publications since the late 1980s (Qian et al., 2020) under the new administration.

This notable research gap can be partly attributed to a lack of data infrastructure covering publications in both Chinese and English, the two most frequently used academic languages in China. The China National Knowledge Infrastructure (CNKI) and its competitors house publications in Chinese, yet the inaccessibility of the data in large batches creates significant barriers to analyzing publications in Chinese at a massive scale. This study fills the gap by using a novel dataset covering projects and publications funded by the National Science Foundation of China (NSFC), one of the most important government research funders in China in the realms of natural sciences and engineering, to provide empirical evidence for multilingual publication patterns of China. This work-in-progress paper reports preliminary results related to the following two research questions:

RQ1: What languages are used in research output funded by NSFC? This question aims to illustrate how different languages are used in STEM research conducted by researchers in China. In particular, we are interested in how this pattern varies across NSFC departments (knowledge domains) and project classes to show how the use of language is situated in broader social contexts.

RQ2: How does the use of languages evolve? Based on the general language usage patterns, we also aim to analyze how these patterns evolved during the investigated period. Even though our sample does not cover projects and publications after the proposal of new national research policies, our results still show the language landscape of STEM research under the previous SCI-oriented research policies in China as a foundation for future studies.

This research bridges a critical gap in multilingual publication practices in China. This work will significantly promote our knowledge of China as a central participant in the global research system, where multiple languages are used to produce and communicate scientific knowledge.

Methods

Data source

This study used a dataset of NSFC-funded projects and research outcomes acquired in August 2021 from the NSFC output portal[1], the official platform for Principal Investigators (PIs) to report progress and the outcomes at the end of NSFC-funded projects. Therefore, our data is considered highly comprehensive and accurate for understanding the landscape of NSFC funding. This research focuses on journal and conference publications as the primary outcome of research projects, accounting for 90.5% of the total outcomes.

This project analyzes projects and publications funded by all the eight *departments* (disciplinary fields) in NSFC, including Mathematics and Physical Sciences (MPS), Chemical Sciences (Chem), Life Sciences (Life), Earth Sciences (Earth), Engineering & Material Science (EMS), Information Sciences (IS), Management Sciences (MS), and Health Sciences

[1] https://kd.nsfc.gov.cn/

(HS). Two notable features in this taxonomy should be noted. First, the *Department of Information Sciences* in NSFC is strongly situated in the knowledge domain of computer science. Second, *Management Sciences* is the house for many social science disciplines within the scope of NSFC, such as economics, public administration, sociology, and library and information science.

Each disciplinary department funds various project types or *project classes* per NSFC. This study selected the following four types: Key Projects (重点项目; *Key*), General Projects (面上项目; *General*), Young Scientist Projects (青年项目; *Young*), and Projects for Less Developed Regions (地区科学基金项目; *Region*). These four project types account for about 95.8% of all projects supported by NSFC in our dataset. In particular, the *Key, Young,* and *General* projects form a "laddered" structure based on PI seniority. A *Key* project typically lasts five years and is usually considered the most prestigious among the four, with most of its PIs being established scholars with prior NSFC project experience. A *General* project usually spans four years and has the broadest set of eligible applicants among the four, where all researchers in universities and research institutions with permanent PI status are eligible to apply. The *Young* project typically lasts three years and is open to scholars under the biological age of 35. Therefore, *Key*, *General*, and *Young* projects usually go to PIs with decreasing levels of seniority.

Additionally, this study focused on projects funded by NSFC starting from 2010 (project funding year), given the fact that no publication in our sample was found to be published before 2010, potentially due to metadata policies by NSFC. The end year of 2015 allows us to ensure that we have complete outcome records for all projects by the time we collected the data. This sample includes 185,465 projects funded by NSFC from 2010 to 2015 and 2,323,443 corresponding publications published mainly between 2010 and 2020.

Publication language identification

We used the Python package *lingua* (version 1.1.3)[2] to evaluate the language of publications based on their titles in our dataset, the only textual elements in the metadata we collected. This package uses *n*-grams of sizes one to five to calculate the Bayesian probability of a text string belonging to a language. In contrast, most existing packages only use *n*-grams up to size 3. This difference makes *lingua* more accurate for shorter texts based on the experiments conducted by the developers. It also makes *lingua* a good fit for our dataset, given that our data only includes minimal textual information for language detection.

We applied this package to our corpus using its default setting to identify all languages used by researchers in China for publications. It identified 53 languages in publications by researchers in China, but English and Chinese are the dominant ones. The top 10 languages by publications account for 99.8% of the total publications, as shown in Table 1.

Table 1. Top 10 languages by the number of publications

Language	*Publications*	*Share in Sample*
English	1,293,155	55.7%
Chinese	792,053	34.1%
Latin	199,038	8.6%
Yoruba	12,425	0.5%
Esperanto	7,483	0.3%
Tagalog	5,360	0.2%

[2] https://pypi.org/project/lingua/

French	2,858	0.1%
Welsh	2,472	0.1%
German	1,911	0.1%
Tsonga	1,506	0.1%
Top 10 languages	2,318,261	99.8%

However, after manually examining publications assigned to other languages in the table (using a random sample of 200 publications for each language), we found that all titles classified into these languages belong to English. This mistake happens because of the particular words used in these titles. For example, the title "Three homoclinic solutions for second-order p-Laplacian differential system" was classified into Latin and "Adsorption and desorption of uranium (VI) by Fe-Mn binary oxide in aqueous solutions" into French. We thus manually re-classified all publications in the other eight languages to English. Moreover, we manually examined two subsamples of 400 randomly selected publications classified into English and Chinese, respectively. We found no classification in the samples. As a result, we used all 2,318,261 publications covered by these ten languages as the final analytical sample of this research.

Results

The patterns of publication languages by field and project class

Our results show an overwhelming dominance of English in the publishing languages by researchers in China in the past decade. Among the 2,318,261 publications by NSFC projects, 1,526,208 (65.8%) are in English, while 792,053 (34.2%) are in Chinese, adding to the evidence that English is the *lingua franca* in scientific research.

We further confirmed the dominant role of English across all disciplinary fields, yet the extent to which it varies (Figure 1, left panel). Our results show that *Chemical Sciences* (83.7%) has the highest share of English publications among the eight, followed by *Mathematics and Physical Sciences* (79.8%). *Management Sciences*, the house for many social science disciplines in NSFC, has the lowest share of English publications (36.6%), consistent with existing evidence (Stockemer & Wigginton, 2019). Our results show that the share of English publications is positively correlated with the seniority of PIs embedded in the project classes (Figure 1, right panel). Aggregating by project class, projects under *Key* classes have more than 75% of English-language publications, whereas *General* and *Young* have lower yet similar numbers.

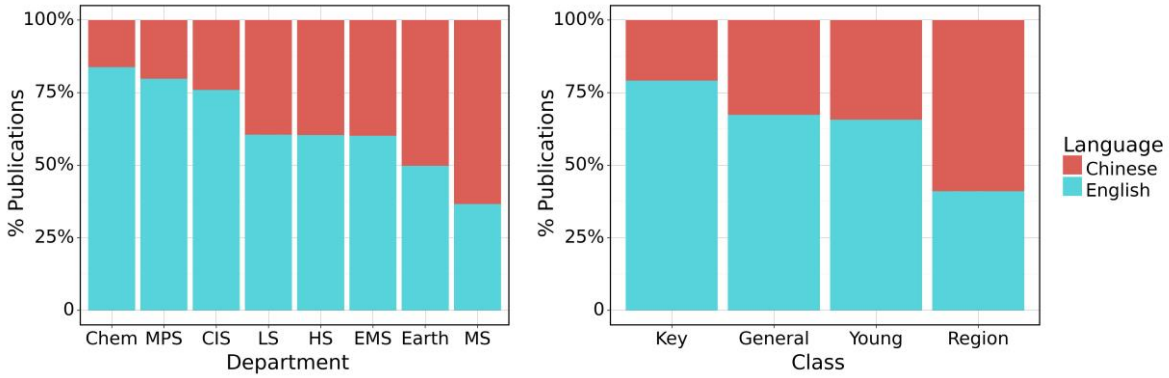

Figure 1. Share of publications in two languages by field (left) and project classes (right)

The temporal trend of publication languages

We further examined the change of publication shares over the examined funding window, i.e., 2010-2015. Our results show a mild decrease in the share of Chinese-language publications over time (Figure 2). It indicates the increasing usage of English in the research outputs by the

research enterprise of China, which might be highly related to the various incentives for publishing in English (especially in SCI-indexed journals). Please note that we examined the issue during the time window before the advent of new research policies in China that stress localization. So the data has not shown any response to the new policy.

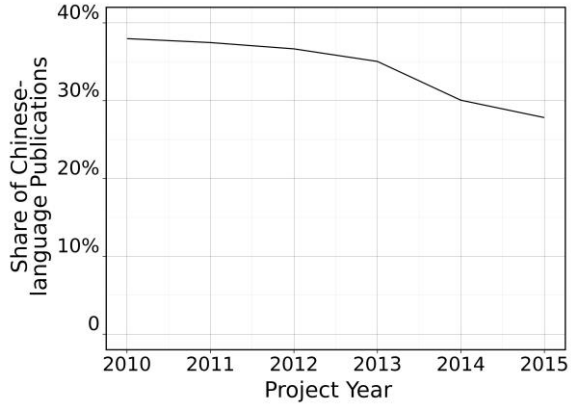

Figure 2. Share of Chinese-language publications over the funding years

We also examined the trend based on disciplinary field and project class (Figure 3). Our results show that all fields and project classes follow a similar trend as the above, indicating the increasing reliance upon English as the leading publishing language by Chinese researchers is a universal phenomenon across fields and researchers of different seniority, although with variance in the extent of change.

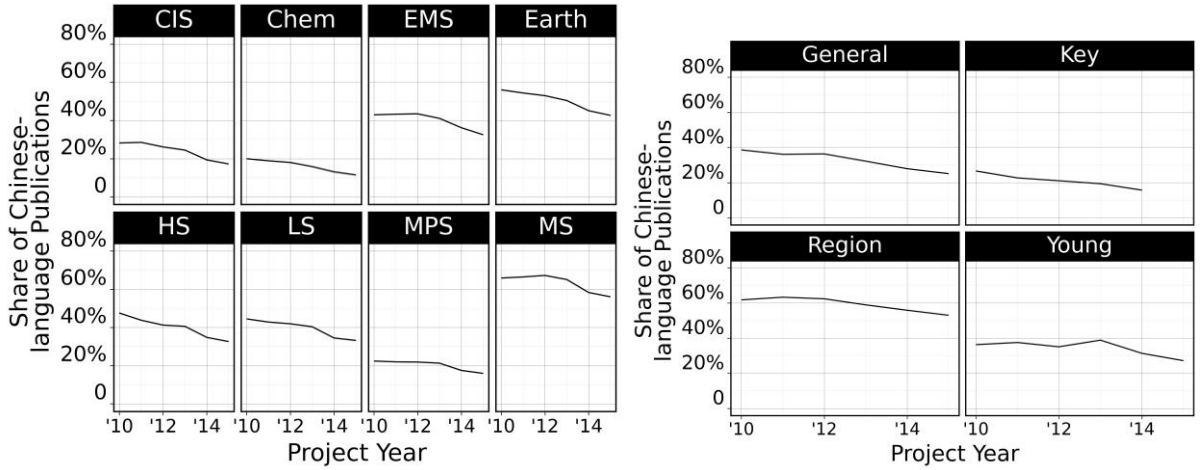

Figure 3. Share of Chinese-language publications over funding years by department (left) and project class (right)

Discussions and conclusions

This work-in-progress paper presents preliminary results concerning using languages for publication by STEM researchers funded by NSFC. This research bridges a critical gap in understanding multilingual publication practices in China, primarily caused by data constraints.

Our preliminary findings confirm that Chinese STEM scholarship heavily relies on the English language to publish research outputs: close to two-thirds of all publications in our sample were published in English. More granular analyses show that hard sciences and engineering more heavily use English to publish than medical and social sciences domains (Stockemer & Wigginton, 2019). Moreover, there is a positive correlation between the amount of English-language publications and the PI seniority of project classes. This finding suggests that projects led by more senior rather than junior researchers are more likely to publish in English, which

warrants more detailed analyses in the future. Our results also show an increasing reliance on English-language publication venues by Chinese STEM researchers during the 2010s. Given that our sample precedes the new research evaluation policies in China, whether the new policies will reverse the trend observed from our sample is an important topic in future research.

As the next step of our project, we will examine how the two languages were used on the project level and particularly what attributes of projects and publications (such as the publication type, PI gender and seniority, and publication topics) might influence the choice of languages for publication. Moreover, we will also aim to understand the use of translation services by Chinese researchers, an important factor underlying the use of multiple academic languages globally. We hope our research will serve as the basis for more comprehensive discussions regarding the multilingual publishing practice and policies in China.

References

González-Alcaide, G., Valderrama-Zurián, J. C., & Aleixandre-Benavent, R. (2012). The Impact Factor in non-English-speaking countries. *Scientometrics*, *92*(2), 297–311. https://doi.org/10.1007/s11192-012-0692-y

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., Guns, R., Nowotniak, R., Petr, M., Sivertsen, G., Istenič Starčič, A., & Zuccala, A. (2018). Publication patterns in the social sciences and humanities: evidence from eight European countries. *Scientometrics*, *116*(1), 463–486. https://doi.org/10.1007/s11192-018-2711-0

Kulczycki, E., Guns, R., Pölönen, J., Engels, T. C. E., Rozkosz, E. A., Zuccala, A. A., Bruun, K., Eskola, O., Starčič, A. I., & Petr, M. (2020). Multilingual publishing in the social sciences and humanities: A seven-country European study. *Journal of the Association for Information Science and Technology*, *71*(11), 1371–1385.

Liang, W., Gu, J., & Nyland, C. (2022). China's new research evaluation policy: Evidence from economics faculty of Elite Chinese universities. *Research Policy*, *51*(1), 104407. https://doi.org/https://doi.org/10.1016/j.respol.2021.104407

Liu, F., Hu, G., Tang, L., & Liu, W. (2018). The penalty of containing more non-English articles. *Scientometrics*, *114*(1), 359–366. https://doi.org/10.1007/s11192-017-2577-6

Mathies, C., Kivistö, J., & Birnbaum, M. (2020). Following the money? Performance-based funding and the changing publication patterns of Finnish academics. *Higher Education*, *79*(1), 21–37. https://doi.org/10.1007/s10734-019-00394-4

Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, *106*(1), 213–228.

Qian, J., Yuan, Z., Li, J., & Zhu, H. (2020). Science Citation Index (SCI) and scientific evaluation system in China. *Humanities and Social Sciences Communications*, *7*(1), 108. https://doi.org/10.1057/s41599-020-00604-w

Sivertsen, G. (2018). Balanced multilingualism in research. *L'évaluation de La Recherche Scientifique: Enjeux, Méthodes et Instruments, Actes de La Colloque International*, 88–102.

Sivertsen, G., & Larsen, B. (2012). Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential. *Scientometrics*, *91*(2), 567–575.

Stockemer, D., & Wigginton, M. J. (2019). Publishing in English or another language: An inclusive study of scholar's language publication preferences in the natural, social and interdisciplinary sciences. *Scientometrics*, *118*(2), 645–652. https://doi.org/10.1007/s11192-018-2987-0

van Leeuwen, T., Moed, H., Tijssen, R., Visser, M., & van Raan, A. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, *51*(1), 335–346.

Zhou, P., Su, X., & Leydesdorff, L. (2010). A comparative study on communication structures of Chinese journals in the social sciences. *Journal of the American Society for Information Science*

and Technology, *61*(7), 1360–1376.