Multilingualism in Scholarly Publishing: How Far Can Technology Take Us and What Else Can We Do?

LYNNE BOWKER
UNIVERSITÉ LAVAL, CANADA
ORCID: 0000-0002-0848-1035

Disparities created by the use of English as the key language for scholarly publishing are becoming increasingly clear in many disciplines. Tatsuya Amano et al. (2023) surveyed environmental scientists around the globe and found that it takes non-native speakers of English substantially more time, effort, and money to read and write articles in English. Jacky Deng and Alison Flynn (2023, 1529) interviewed non-Anglophone graduate students in chemistry and learned that for many, communicating research in English is "their most pervasive challenge." In the field of digital humanities (DH), Puthiya Purayil Sneha (2022, 15) emphasizes that "the prevalent global discourse around DH is largely Anglocentric," while Roopika Risam (2018, 79) points out that this often leads to "centering epistemologies and ontologies of the Global North, namely the U.S. and western Europe, which in turn decenters those of Indigenous communities and the Global South." Scholars who publish in languages other than English are cited less often (Di Bitetti and Ferreras 2017), and there is "a persistent lack of international representation on editorial boards" (Espin et al. 2017). But while the problems stemming from the use of a single language for science are becoming ever clearer, the path forward is less obvious.

For instance, if all scholars publish in their own language, how will others evaluate, discover, or read their work? Some are pinning their hopes on technologies, such as automatic translation tools (e.g., Google Translate) and tools based on large language models (LLMs) (e.g., ChatGPT) that are becoming increasingly prevalent. In principle, such tools could support the use of multiple languages in the scholarly communication ecosystem. Imagine a scenario where an author from Chile submits a manuscript to a journal in Spanish. The editor identifies a subject expert in Japan, who uses a translation tool to get a version in Japanese and then prepares their peer review feedback in Japanese. This goes back to the editor, who machine translates the feedback into Spanish for the author. Following revisions, the article is published in Spanish, but scholars in Greece, Egypt, Thailand, or elsewhere can in turn use translation tools to read the

https://doi.org/10.3998/jep.6262

article in their own language. This scenario describes a truly multilingual environment, where various actors in the scholarly communication ecosystem can undertake their activities in a language of their own choosing. In principle, it could work, but in practice, we are not quite there yet.

Data-driven tools and their implications for less widely used languages

Tools such as Google Translate and ChatGPT use data-driven approaches, such as machine learning. For a task such as translation, the data consist of previously translated texts. This means gathering a corpus of texts in one language and their translations into another language. Preferably, these translations have been done by professional translators so that the computer will learn from good examples rather than from crummy ones. However, computers cannot learn from just a few examples. For machine learning tasks, computers need millions, or sometimes even billions, of examples (Pérez-Ortiz, Forcada, and Sánchez-Martínez 2022). For languages that are widely used and for which there is a lot of translation activity, it is relatively easy to find examples of previously translated texts. Taking English and French as examples, these are languages used in many countries around the world, and there is a lot of translation activity between them. In Canada alone, the federal government's Translation Bureau translated 343 million words for governmental departments, agencies, and Parliament in 2019 to 2020 (Government of Canada 2021), while the City of Ottawa's French language services translated more than 13 million words in 2021 (Willing 2022). Of course, translation between these two languages also takes place in other regions of Canada, in other bilingual countries (e.g., Cameroon), in the various bodies of the European Union, at organizations such as the United Nations and the World Health Organization, and more. Therefore, it is easy to see how tool developers could compile a suitably large training corpus for translation between English and French. In contrast, other languages have far fewer speakers (e.g., Cree, Welsh). Moreover, even if two languages independently have a substantial number of speakers (e.g., Russian and Urdu), there may not be a significant volume of translation activity between them, meaning that it would be hard to create a large training corpus of translated texts. Languages and language pairs for which many texts and translations are available are described as high resource, while those for which fewer texts and translations can be located are known as low resource. This concept of high- and low-resource languages can also apply to domains and text types. For instance, some subjects are very common, while others—like many research topics—are more specialized. In order for a computer to learn relevant terminology or a particular style, those features need to be included in sufficient numbers in the training corpus. This

shrinks the pool of resources even more. Good luck finding millions of examples of texts translated between Tamil and Czech on hydrogeology! Another feature of these tools is that they tend to produce better quality translations between languages that are closely related. The more distant the languages, the poorer the translation quality.

So what does this mean when it comes to using translation tools for scholarly publishing? Essentially, it means that these tools will do a better job of supporting researchers who work with widely used languages, such as English and French, while those researchers who work with less common languages risk being poorly served or not served at all. Google Translate currently supports 133 languages, while Meta AI's ambitious No Language Left Behind (NLLB) project (Costa-jussà et al. 2022) aims to support 200. These efforts are moving us in the right direction, but they are far from meeting the needs of all the speakers of the world's more than 7,000 languages. Even so, translation tools could potentially help to inject a greater degree of linguistic diversity into scholarly publishing, if only for some of the more widely spoken languages. But are they doing so?

Supporting change or reinforcing bad habits?

Early evidence suggests that the most common use of language technologies in scholarly publishing is *not* to support multilingual publishing. For example, some journal editors (e.g., Thorp 2023) have sought to rule out the use of generative AI tools in any stage of manuscript production. In response, Mohamed Seghier (2023) and other scholars for whom English is an additional language counter that using free tools such as ChatGPT or DeepL Translator to help with editing and proofreading in English should be permitted. Violeta Berdejo-Espinola and Tatsuya Amano (2023, 991) go so far as to argue that "reducing the technical and financial burden of editing and proofreading papers for nonnative English speakers would be a substantial step toward achieving equity in science." Indeed, in a survey of 1,600 researchers, Richard Van Noorden and Jeffrey Perkel (2023) asked respondents to identify what they saw as the biggest benefit of generative AI for research. For more than half of the respondents, "the clearest benefit . . . was that LLMs aided researchers whose first language is not English, by helping to improve the grammar and style of their research papers, or to summarize or translate other work" (Van Noorden and Perkel 2023, 674).

For their part, Mohammad Hosseini and Serge P. J. M. Horbach (2023) suggest that using LLMs could help editors to overcome reviewer shortages. Because these tools can support peer reviewers with the task of preparing their reports, Hosseini and Horbach suggest that editors could access a larger and more efficient pool of candidate reviewers: "LLMs can also increase the pool of reviewers by opening it up to non-native English speakers (some of whom might be able to use various translation services to

read a paper) and feed their opinion/views in broken English to LLMs and ask them to write a more presentable review in English" (4).

Meanwhile, Lynne Bowker, Philips Ayeni, and Emanuel Kulczycki (2023) conducted a systematic review of the literature at the intersection of language technologies and scholarly communication. In all 40 of the studies included their review, English features as one of the languages in the translation pair, and just one-quarter of the studies include a low-resource language. Moreover, in nearly two-thirds of the studies, English is the *target* language, and the focus is on non-Anglophone scholars using translation tools as writing aids to produce texts for publication in English.

What is striking in all of these examples is that they suggest that translation tools are not necessarily helping to displace English as the key language of scholarly communication. Instead, non-Anglophone scholars are using these tools to reduce the burden of preparing English-language publications, but this is not necessarily creating a genuinely multilingual scholarly communication ecosystem. Instead, the responsibility for translation in scholarly publishing continues to rest on the shoulders of non-Anglophone scholars, while the English language and English-speaking scholars remain in a privileged position.

Other tools and resources to support multilingualism in scholarly publishing

What can we learn from this? One key observation is that technology alone is not enough to achieve or sustain a multilingual scholarly communication ecosystem. Current translation tools, while not perfect, can carry out translation in multiple directions (at least for high-resource languages) and can support tasks such as discovering and reading research that has been written in other languages. Yet these tools are mainly being used to translate out of other languages and into English, thus reinforcing rather than diversifying this largely monolingual ecosystem. Certainly more research is needed to find techniques for better supporting low-resource languages, but there is also a need for policies that shift the responsibility away from expecting speakers of other languages to use translation tools to produce texts in English and towards encouraging the use of these tools to access and engage with research that has been written in other languages.

Beyond technology, what can different actors in scholarly publishing do to better support multilingualism? The Helsinki Initiative on Multilingualism in Scholarly Communication (2019) undertakes advocacy work to encourage policy makers to value and support research in multiple languages. They raise awareness through their seminar series and other activities. Likewise, the UNESCO (2021) Recommendation on Open Science considers multilingualism to be one facet of openness.

Various researchers have begun to share recommendations based on their own lived experiences (e.g., Khelifa, Amano, and Nuñez 2022; Nolde-Lopez et al. 2023; Steigerwald et al. 2022), but these are currently scattered in the literature belonging to different disciplines. Bringing this information together in a single open resource is one of the goals of the Developing Institutional Open Access Publishing Models to Advance Scholarly Communication (DIAMAS) project. DIAMAS is developing an Extensible Quality Standard for Institutional Publishing (EQSIP) for Diamond Open Access, and multilingualism is a key element addressed in the equity, diversity, inclusion, and belonging (EDIB) component of the EQSIP. Team members have combed the literature and compiled an open toolsuite of recommendations for ways in which authors/researchers, peer reviewers, editors and editorial board members, librarians, and journal and book publishers can facilitate multilingualism in scholarly publishing (Bowker et al., 2024). Here is a small selection of recommendations from the toolsuite, where they are accompanied by links to examples and additional information:

- **For authors:** Practice citation diversity (e.g., citing research published in other languages) and consider including a citation diversity statement with your own articles, including any linguistic limits on literature searches (e.g., searches that have been conducted only in English).
- **For peer reviewers:** Identify to the editors the languages in which you are able to provide peer review feedback.
- **For editors:** Provide/recommend guidelines to help authors prepare manuscripts in a reader- and (machine) translation-friendly way. Well-crafted input can lead to better quality translation output from automatic translation tools.
- **For librarians:** Add multilingual metadata to items to facilitate multilingual searches in library catalogs.
- **For journal publishers:** Translate abstracts, summaries, and tables of contents into multiple languages.

Another group, Open Scholarly Communication in the European Research Area for Social Sciences and Humanities (OPERAS), has the mission to coordinate and federate resources in Europe to efficiently address scholarly communication needs. OPERAS has a special interest group on multilingualism, which has already contributed to the development of a multilingual platform called GoTriple to support search and discovery in nine languages (with more to be added). In addition, OPERAS is currently exploring the development of a scientific translation service combining open source technological tools and resources with human skills to support the translation process within scholarly publishing (Fiorini et al. 2020; Fiorini 2023). A similar proposal has been put forward by the French language commissioner of Quebec in Canada to better

support the country's French language researchers (Dubreuil, Tremblay-Faulkner, and Parent 2023).

Finally, the global Coalition for Advancing Research Assessment (CoARA) has set out a shared direction for changes in research assessment practices intended to maximize the quality and impact of research (CoARA 2022). To this end, CoARA has established a Working Group on Multilingualism and Language Biases in Research Assessment, which has the dual objectives of raising awareness about the importance of multilingualism in scholarly publishing and providing guidelines for recognizing, rewarding and incentivizing research published in all languages (Donahoe 2024).

Overall then, the picture is encouraging. As demonstrated by this special issue, along with other initiatives, there is clearly an appetite for a more multilingual scholarly communication ecosystem, which can include but not rely solely on technology to diversify its linguistic base. Although the way forward is not yet entirely clear, this challenge cuts across nearly all academic disciplines, so it is important to break down the silos and to learn from and support one another as we strive for improved linguistic equity in research. Let's keep the momentum going.

References

Amano, Tatsuya, Valeria Ramírez-Castañeda, Violeta Berdejo-Espinola, Israel Borokini, Shawan Chowdhury, Marina Golivets, Juan David González-Trujillo, et al. 2023. "The Manifold Costs of Being a Non-native English Speaker in Science." *PLoS Biology* 21 (7): e3002184. https://doi.org/10.1371/journal.pbio.3002184.

Berdejo-Espinola, Violeta, and Tatsuya Amano. 2023. "AI Tools Can Improve Equity in Science." *Science* 379 (6636): 991. https://doi.org/10.1126/science.adg9714.

Bowker, Lynne, Philips Ayeni, and Emanuel Kulczycki. 2023. *Linguistic Privilege and Marginalization in Scholarly Communication: Understanding the Role of New Language Technologies for Shifting Language Dynamics*. Final report submitted to the Social Sciences and Humanities Research Council of Canada, December 17, 2023. https://doi.org/10.20381/858s-q632.

Bowker, Lynne, Mikael Laakso, Janne Pölönen, and Claire Redhead. 2024. "What Can Be Done about Scholarly Communication's Diversity Problem?" *LSE Impact Blog*, May 23, 2024. Accessed June 18, 2024. https://blogs.lse.ac.uk/impactofsocialsciences/2024/05/23/what-can-be-done-about-scholarly-communications-diversity-problem/.

Coalition for Advancing Research Assessment (CoARA). 2022. *Agreement on Reforming Research Assessment*. Accessed June 18, 2024. https://coara.eu/agreement/the-agreement-full-text/.

———. n.d. Accessed June 18, 2024. https://coara.eu.

Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, et al. 2022. *No Language Left Behind: Scaling Human-Centered Machine Translation*. https://arxiv.org/abs/2207.04672.

Deng, Jacky M., and Alison B. Flynn. 2023. " 'I Am Working 24/7, but I Can't Translate That to You': The Barriers, Strategies, and Needed Supports Reported by Chemistry Trainees from English-as-an-Additional Language Backgrounds." *Journal of Chemical Education* 100 (4): 1523–36. https://doi.org/10.1021/acs.jchemed.2c01063.

Di Bitetti, Mario S., and Julián A. Ferreras. 2017. "Publish (in English) or Perish: The Effect on Citation Rate of Using Languages Other than English in Scientific Publications." *Ambio 46* (1): 121–27. https://doi.org/10.1007/s13280-016-0820-7.

DIAMAS (Developing Institutional Open Access Publishing Models to Advance Scholarly Communication). n.d. Accessed January 12, 2024. https://diamasproject.eu.

Donahoe, Casey. 2024. "Multilingualism and Language Bias in Research Assessment: Supporting Non-English Speaking Scientists." Declaration on Research Assessment (DORA), January 18, 2024. Accessed June 18, 2024. https://sfdora.org/2024/01/18/multilingualism-and-language-bias-in-research-assessment-supporting-non-english-speaking-scientists/.

Dubreuil, Benoit, Marc Tremblay-Faulkner, and Rodolphe Parent. 2023. *Le français, langue du savoir?* Québec: Commissaire à la langue française.

Espin, Johanna, Sebastian Palmas, Farah Carrasco-Rueda, Kristina Riemer, Pablo E. Allen, Nathan Berkebile, Kirsten A. Hecht, et al. 2017. "A Persistent Lack of International Representation on Editorial Boards in Environmental Biology." *PLoS Biology* 15 (12): e2002760. https://doi.org/10.1371/journal.pbio.2002760.

Fiorini, Susanna. 2023. *Exploratory Studies for the Creation of a Technology-Aided Collaborative Translation Service in Open Scholarly Communication: General Report*. https://zenodo.org/records/10972986.

Fiorini, Susanna, Franck Barbin, Martine Garnier-Rizet, Katell Hernandez Morin, Franziska Humphreys, Amélie Josselin-Leray, Natalie Kübler, et al. 2020. *Rapport du groupe de travail "Traductions et science ouverte*." Comité pour la science ouverte. https://doi.org/10.52949/20.

Google Translate. n.d. Accessed January 12, 2024. https://translate.google.com.

Government of Canada. 2021. "Translation Bureau: Standing Committee on Official Languages—February 16, 2021." Standing Committee on Official Languages. https://www.tpsgc-pwgsc.gc.ca/trans/documentinfo-briefingmaterial/lang/2021-02-16/p1-eng.html.

Helsinki Initiative. 2019. *Helsinki Initiative on Multilingualism in Scholarly Communication*. Helsinki: Federation of Finnish Learned Societies, Committee for Public Information, Finnish Association for Scholarly Publishing, Universities Norway, and European Network for Research Evaluation in the Social Sciences and the Humanities. https://doi.org/10.6084/m9.figshare.7887059.

Hosseini, Mohammad, and Serge P. J. M. Horbach. 2023. "Fighting Reviewer Fatigue or Amplifying Bias? Considerations and Recommendations for Use of ChatGPT and Other Large Language Models in Scholarly Peer Review." *Research Integrity and Peer Review* 8 (1): Article 4. https://doi.org/10.1186/s41073-023-00133-5.

Khelifa, Rassim, Tatsuya Amano, and Martin A. Nuñez. 2022. "A Solution for Breaking the Language Barrier." *Trends in Ecology and Evolution* 37 (2): 109–12. https://doi.org/10.1016/j.tree.2021.11.003.

Nolde-Lopez, Bianca, Joanna Bundus, Henry Arenas-Castro, Diego Román, Shawan Chowdhury, Tatsuya Amano, Violeta Berdejo-Espinola, and Susana M. Wadgymar. 2023. "Language Barriers in Organismal Biology: What Can Journals Do Better?" *Integrative Organismal Biology* 5 (1): 1–15. https://doi.org/10.1093/iob/obad003.

OPERAS (Open Scholarly Communication in the European Research Area for Social Sciences and Humanities). n.d. Accessed January 12, 2024. https://operas-eu.org.

Pérez-Ortiz, Juan Antonio, Mikel L. Forcada, and Felipe Sánchez-Martínez. 2022. "How Neural Machine Translation Work." In *Machine Translation for Everyone: Empowering Users in the Age of Artificial Intelligence*, edited by Dorothy Kenny, 141–64. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.6760020.

Risam, Roopika. 2018. "Decolonizing the Digital Humanities in Theory and Practice." In *Routledge Companion to Media Studies and Digital Humanities*, edited by Jentery Sayers, 78–86. New York: Routledge.

Seghier, Mohamed L. 2023. "Using ChatGPT and Other AI-assisted Tools to Improve Manuscripts Readability and Language." *International Journal of Imaging Systems and Technology* 33 (3): 773–75. https://doi.org/10.1002/ima.22902.

Sneha, Puthiya Purayil. 2022. "Alternative Histories of Digital Humanities: Tracing the Archival Turn." In *Global Debates in the Digital Humanities*, edited by Domenico Fiormonte, Sukanta Chaudhuri, and Paola Ricaurte, 15–27. Minneapolis: University of Minnesota Press.

Steigerwald, Emma, Valeria Ramírez-Castañeda, Débora Y. C. Brandt, András Báldi, Julie Teresa Shapiro, Lynne Bowker, and Rebecca D. Tarvin. 2022. "Overcoming Language Barriers in Academia: Machine Translation Tools and a Vision for a Multilingual Future." *BioScience* 72 (10): 988–98. https://doi.org/10.1093/biosci/biac062.

Thorp, H. Holden. 2023. "ChatGPT Is Fun, but Not an Author." *Science* 379 (6630): 313. https://www.science.org/doi/10.1126/science.adg7879.

UNESCO. 2021. Recommendation on Open Science. https://en.unesco.org/science-sustainable-future/open-science/recommendation.

Van Noorden, Richard, and Jeffrey M. Perkel. 2023. "AI and Science: What 1,600 Researchers Think." *Nature* 621:672–75.

Willing, Jon. 2022. "Vaccine Clinics Generate French-Language Complaints." *Ottawa Citizen*, February 25, 2022. https://ottawacitizen.com/news/local-news/vaccine-clinics-generate-french-language-complaints.