International Journal of Infectious Diseases 96 (2020) 352–354

Contents lists available at ScienceDirect

International Journal of Infectious Diseases

journal homepage: www.elsevier.com/locate/ijid

Short Communication

# On the crucial role of multilingual biomedical databases in epidemic events (SARS-CoV-2 analysis)

Felipe Soares[a,b,*], Gabrielli Harumi Yamashita[b]

[a] University of Sheffield, Computer Science Department, NLP Group, United Kingdom
[b] Universidade Federal do Rio Grande do Sul, Departamento de Engenharia de Produção, Brazil

ARTICLE INFO

*Article history:*
Received 16 March 2020
Received in revised form 4 May 2020
Accepted 7 May 2020

*Keywords:*
Epidemics
Natural Language Processing
Biomedical databases

ABSTRACT

The need for multilingual biomedical databases was already pointed out by different authors. They argue about the need for making translations available in other languages and centralized access to regional databases and that one should not disregard citations in other languages. This fact could not be any more real in the current situation regarding the novel coronavirus. When considering treatment, diagnosis and prevention, around 44% of the articles in PubMed were written in Chinese. This prompts the urgent need for quality automatic translation to make such extremely valuable information available to medical personnel in as many languages as possible. We also point out that the community should also make efforts to guarantee editorial quality and to follow the best practices in editing and publishing. This is of critical importance as well, such that the content is properly scrutinized before being published.



The need for multilingual biomedical databases has already been pointed out by Prieto (2018). He argues the need for making translations available in other languages and for centralized access to regional databases. Lazarev and Nazarovets (2018) point out the fact that one should not disregard citations in other languages. This fact could not be any more real in the current situation regarding the novel coronavirus. In addition, a fair and thorough review process should guarantee the final required quality for decision-makers. As an example, we have recently seen a dubious article published (in English) praising the use of chloroquine to treat COVID-19 patients. However, the study had serious design flaws that were overlooked by reviewers that practically led to mistaken or exaggerated conclusions. Ultimately, recent appropriate clinical trials are finding divergent evidences (while following a rigorous protocol).

In a search on PubMed/Medline, as of 11/Feb/2020 (when the disease was still confined to China), it was possible to retrieve 84 articles regarding the novel coronavirus, 74 of them being in English. However, most of them are related to the genomics of the virus or epidemiologic analysis. Those studies are extremely important, but are they relevant to the physicians that are on the front line of this infectious event?

When considering treatment, diagnosis and prevention, around 44% of the articles were written in Chinese. This prompts the urgent need for high-quality automatic translation to make such extremely valuable information available to medical personnel in as many languages as possible. This is to some extent already under active development, with the WMT Conference (Conference on Machine Translation) being the main venue with a specific track for biomedical articles. In the two most recent WMT conferences (2018 and 2019)[1] interesting results were reported for the English/Portuguese and English/Spanish language pairs. For instance, for the English to Spanish, the number of automatic translated sentences judged by humans as better than human translations was larger than the number of human sentences judged better than the automatic ones. When combining the number of times that the best automatic translation was equally good or better than human translation for WMT19, we get an average of 73% of correct translations according to human judgment, with a surprising 90% for EN/ES (English/Spanish) and 82.09% for ZH/EN (Chinese/English). This strengthens our point that automatic translation can indeed be used to aid dissemination of biomedical scientific content.

* Corresponding author at: University of Sheffield, Computer Science Department, NLP Group, United Kingdom.
*E-mail address:* fs@felipesoares.net (F. Soares).

[1] http://www.statmt.org/wmt19/ and http://www.statmt.org/wmt18/.

https://doi.org/10.1016/j.ijid.2020.05.023

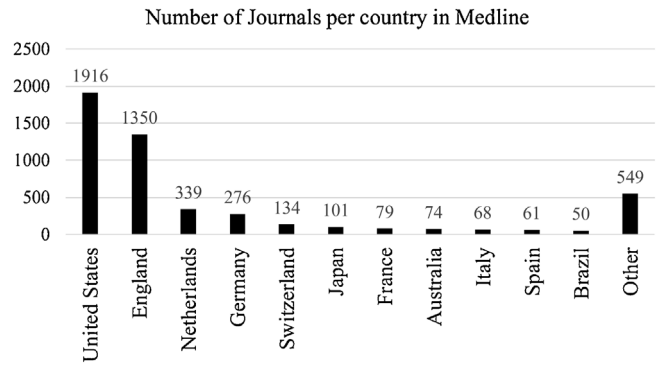


**Figure 1.** Journals per country indexed in Medline, showing countries with at least 50 indexed journals.

As another example of the importance of non-English articles, we checked the Chinese Medical Full-text Database Yiigle[2] for the novel coronavirus. It is possible to retrieve 13 additional articles with 11 of them (85%) related to medical staff protection against the virus, rehabilitation guidelines, possible treatments, and sterilization protocols. Many of these topics are not present in PubMed/Medline, which contains predominantly articles in English. However, it is exactly this particular kind of information that can help physicians to fight this infection, and they should be provided access to the most up-to-date information to make informed decisions. Thus, these articles should be available in as many languages as possible, through the help of automatic translation, for instance.

As the main biomedical database, which should encompass as many languages as possible, PubMed/Medline states that it already indexes articles in other languages, provided an abstract in English is available. However, when querying the journals indexed in Medline, only 57 out of 4997 are in another language that is not English. In addition, we also checked the countries of publication, which could give an idea of Medline's geographical coverage and are shown in Figure 1. The prevalence of American and English journals is noticeable, as well as from developed countries, with Brazil as the only developing country with at least 50 indexed journals. Thus, many other portals were proposed to fill this gap, such as LILACS (Latin-American and Caribbean countries), Yiigle (China), INDMED (India), UDB-MED (Russia). Each portal usually has its own specific set of rules for indexing.

For inclusion in PubMed/Medline, journals need to fulfil a list of criteria, including explicit external peer-reviewing and adherence to ethical guidelines. Thus, being indexed in Medline is usually seen as a quality indicator. However, regional databases may not always follow such quality criteria. For instance, the Yiigle database does not explicitly state their inclusion criteria, thus one cannot be certain about peer-reviewing, for instance. The LILACs database, on the other hand, has a very similar selection and permanence criteria to Medline,[3] including peer-reviewing and geographical concentration of the editorial board. Thus, we can see that inclusion criteria, and consequently perceived quality, is not homogeneous between databases and translation alone may not be the only "obstacle" for useful information availability for healthcare professionals' decision-making.

The community should also make efforts to guarantee editorial quality and to follow the best practices in editing and publishing. This is of critical importance as well, such that the content is properly scrutinized before being published. As an example, we can draw special attention to the peer-review step, which can be troublesome in regional journals. For instance, some Brazilian journals indexed in LILACs and Medline may carry the whole reviewing process in Portuguese, then translating the article when it is approved. Thus, peer-review is done predominantly by researchers of Portuguese-speaking countries, which can introduce bias. Another concerning example is regarding tightly controlled, or under embargo, countries, such as China or Iran, where researchers may not have easy access to international content (Normile, 2017; Saeidnia and Abdollahi, 2013), also introducing a critical bias.

Another example of possible peer-review bias is when reviewers are selected mainly from one institution or more than one reviewer is from the same research group. This could bias the peer-review since people from the same research group/laboratory probably share the same opinions, and they should be "weighted" accordingly. Some more "subjective" biases may be related to affiliation, when an author is from a prestigious research group or university, the review may be less scrutinized. A good example of bias reduction is the computer science field, which has greatly improved in the past years, with submissions to most of its conferences being double-blind reviewed, and sometimes the complete peer-review process is published along with the paper. The Open Review[4] platform is steadily increasing the number of conferences covered.

As a concrete example and suggestion, there could be guidelines similar to the International Committee of Medical Journal Editors (ICMJE) regarding authorship. We could (and should) have a strict guideline for selecting reviewers. Some journals, as explained before, have already taken measures to comply with PubMed/Medline, for instance. Some criteria usually are related to avoiding reviewers from the same institution, aiming for geographical and institutional diversity among reviewers. However, we advocate that these criteria should be better elaborated and standardized, such as ICMJE's authorship contribution.

Thus, we ask the scientific community, especially those working in the field of biomedical natural language processing (NLP) and automatic translation, to also devote their valuable efforts in working with other languages rather than only English, and to join efforts to remove language as a barrier in science. But making scientific articles alone more accessible is not enough for health-care professionals to make informed decisions, we also need to ensure that this information has been well (and fairly) scrutinized. Therefore, editorial boards of journals and scientific databases should also seek good practices in guaranteeing good quality editorial work, for instance by ensuring a heterogeneous peer-reviewing process (reviewers from different countries and institutions) and transparent rules for such selection. Members of the research community could also engage in the discussion of creating such a set of recommendations for peer-reviewing and making them official.

## Ethical approval

The authors state that no ethical approval was needed for this manuscript.

[2] http://journal.yiigle.com/Paper/Search?q=%222019%E6%96%B0%E5%9E%8B%E5%86%A0%E7%8A%B6%E7%97%85%E6%AF%92%22&viewBy=pro&sort=ArtPubDate+desc&n=20.

[3] https://lilacs.bvsalud.org/en/lilacs-journal-selection-and-permanence-criteria-2010/.

[4] https://openreview.net/.

## Conflict of interest

The authors state no conflict of interest.

## Acknowledgments

We would like to acknowledge the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) for the provided funding.

## References

Lazarev VS, Nazarovets SA. Don't dismiss citations to journals not published in English [Review of Don't dismiss citations to journals not published in English]. Nature 2018;556(7700):174.

Normile D. Science suffers as China's internet censors plug holes in Great Firewall. 2017 Retrieved from https://www.sciencemag.org/news/2017/08/science-suffers-china-s-internet-censors-plug-holes-great-firewall.

Prieto D. Make research-paper databases multilingual. Nature 2018;560(7716):29.

Saeidnia S, Abdollahi M. Consequences of international sanctions on Iranian scientists and the basis of science. Hepat Month 2013;13(9).