Linguistic Diversity and Representation Gaps in COVID-19 Research: A Systematic Analysis of the CORD-19 Dataset

Charles Watson Ndethi Kibaki

2025-04-06

Abstract

The COVID-19 pandemic generated an unprecedented volume of scientific literature, yet access to this critical information may be limited by language barriers. This research proposal outlines a systematic investigation of linguistic diversity within the COVID-19 Open Research Dataset (CORD-19), with particular focus on identifying representation gaps for low-resource languages. The study will analyze language distribution patterns, assess content availability across languages, and evaluate text complexity and named entity distributions to identify specific disparities. Using language identification techniques and content analysis methods, this research aims to quantify the extent to which low-resource languages, particularly African languages, are underrepresented in COVID-19 scientific literature. The findings will provide empirical evidence of information access inequities and create a foundation for future work in cross-lingual information access, particularly for underrepresented language communities. This study has implications for global health information equity and will inform strategies for extending scientific knowledge to low-resource language speakers.

0.1 1. Introduction

0.1.1 1.1 Background

The COVID-19 pandemic triggered an explosion of scientific research as the global community raced to understand and combat the novel coronavirus. This unprecedented situation led to the creation of the COVID-19 Open Research Dataset (CORD-19), a comprehensive collection of scientific papers about COVID-19, SARS-CoV-2, and related coronaviruses. CORD-19 represents a significant effort to consolidate scientific knowledge, with over 1,000,000 scholarly articles, including more than 400,000 with full text (**wang2020cord19?**).

While this resource has been invaluable for research and policy development, questions remain about its linguistic diversity and the extent to which critical information is accessible across different language communities. Language barriers in health information access have been identified as significant factors contributing to health disparities (**pazzanese2020?**). These disparities may be particularly acute for speakers of low-resource languages – languages with limited digital presence, computational resources, and NLP tools (Joshi et al., 2020).

The global nature of the COVID-19 pandemic requires global information sharing. However, the predominance of English in scientific publishing may create significant information gaps for non-English speaking populations, especially those speaking low-resource languages. Understanding these gaps is crucial for developing strategies to enhance information equity in global health crises.

0.1.2 1.2 Problem Statement

Despite the comprehensive nature of the CORD-19 dataset, there is limited understanding of its linguistic diversity and the extent to which low-resource languages are represented. This creates potential disparities in information access, which may contribute to health inequities and hamper global pandemic re-

sponse efforts. Additionally, without empirical evidence of these representation gaps, it is difficult to prioritize translation efforts or develop appropriate cross-lingual information retrieval systems.

This research seeks to address this gap by systematically analyzing the linguistic diversity of the CORD-19 dataset, with particular attention to the representation of low-resource languages, including African languages. By identifying specific patterns of language distribution and representation gaps, this study will provide crucial groundwork for future efforts to improve cross-lingual information access in scientific domains.

0.1.3 1.3 Research Question and Objectives

This study is guided by the primary research question:

What patterns of linguistic diversity exist within the CORD-19 dataset, and what representation gaps can be identified for low-resource languages?

To address this question comprehensively, the research has the following specific objectives:

1. Quantify the distribution of languages across the CORD-19 dataset using automated language identification methods.

2. Analyze the relationship between language availability and content characteristics such as publication date, topic area, and document type.

3. Assess text complexity and readability metrics across languages to identify potential barriers to information accessibility.

4. Examine named entity distributions across languages to identify terminology gaps that may affect health information understanding.

5. Identify specific representation gaps for low-resource languages, with particular focus on African languages.

0.1.4 1.4 Significance of the Study

This research addresses a critical gap in understanding how scientific information about COVID-19 has been distributed across language communities. The findings will:

- Provide quantitative evidence of language representation gaps in COVID-19 scientific literature
- Identify specific barriers to information access for speakers of low-resource languages
- Inform prioritization strategies for translation efforts and resource development
- Create a foundation for developing more effective cross-lingual information systems
- Contribute to broader discussions about information equity in global health crises

For low-resource African languages specifically, this research will establish baseline measures of representation in scientific literature that can inform future resource development efforts. The methodological approach developed can also be extended to other domains and datasets to assess linguistic diversity and representation gaps more broadly.

0.2 2. Literature Review

0.2.1 2.1 Linguistic Diversity in Scientific Publishing

Scientific publishing has historically been dominated by a small number of languages, with English emerging as the predominant language of international scientific communication (**amano2016?**). This dominance creates challenges for non-English speakers in accessing scientific information and contributes to what has been termed "linguistic imperialism" in academic publishing (**phillipson1992?**).

Several studies have examined linguistic diversity in scientific literature across various disciplines. (**amano2016?**) found that over 75% of biodiversity conservation literature is published in English, creating significant barriers for conservation efforts in non-English speaking regions. Similarly, (**liu2017?**) analyzed the language distribution of medical literature and found that English-language publications

significantly outnumbered those in other languages, even for research conducted in non-English speaking countries.

In the context of COVID-19, (**taskin2020?**) conducted a preliminary analysis of early pandemic literature and found that English-language publications dominated the discourse, potentially limiting information access for non-English speaking healthcare providers and populations. However, there has been limited systematic analysis of language distribution in comprehensive COVID-19 literature collections like CORD-19.

0.2.2 2.2 Low-Resource Languages and NLP

Low-resource languages (LRLs) are typically defined as languages with limited availability of digital resources, computational tools, and NLP systems (**cieri2016?**). These languages face significant challenges in the digital age, as they often lack basic resources such as digital corpora, part-of-speech taggers, or machine translation systems (Joshi et al., 2020).

Joshi et al. (2020) proposed a taxonomy of languages based on their resource availability, categorizing them from "The Winners" (languages with abundant resources) to "The Left-Behinds" (languages with virtually no digital resources). According to their analysis, the majority of the world's languages fall into the lower categories, highlighting substantial disparities in language technology development.

African languages, in particular, face significant resource challenges. (**nekoto2020?**) documented the state of NLP for African languages, highlighting that most fall into the "Left-Behind" or "Scraping By" categories in Joshi's taxonomy. Their participatory research approach with the Masakhane community demonstrates the potential for community-driven efforts to address these resource gaps.

The limited availability of resources for low-resource languages has direct implications for information access. As (**besacier2014?**) note, speakers of these languages often face a "digital language divide" that restricts their ability to access and contribute to digital content, including critical health information.

0.2.3 2.3 Health Information Disparities

Language barriers have been identified as significant contributors to health disparities globally (**yeheskel2019?**). Limited English proficiency has been associated with poorer health outcomes, reduced access to preventive services, and decreased satisfaction with healthcare (**pazzanese2020?**).

During the COVID-19 pandemic, these language-based disparities became particularly evident. (**piller2020?**) documented how language barriers affected various aspects of pandemic response, from understanding public health directives to accessing testing and treatment information. They found that multilingual information provision was often inadequate, with information in languages other than the dominant national language frequently delayed, reduced in content, or entirely absent.

(**kim2020?**) specifically examined COVID-19 information availability for limited English proficiency populations in the United States and found significant gaps in translated materials, especially for speakers of less common languages. Similar patterns have been observed globally, with (**yeheskel2019?**) noting particular challenges for speakers of indigenous and minority languages.

These studies highlight the real-world impact of language representation gaps in health information. However, most have focused on public health communications rather than scientific literature, leaving a gap in understanding how language barriers might affect access to the growing body of COVID-19 research.

0.2.4 2.4 Cross-Lingual Information Retrieval and Access

Cross-lingual information retrieval (CLIR) systems aim to bridge language gaps by allowing users to find information in languages different from their query language (**zhou2012?**). These systems have become increasingly sophisticated with advances in neural

machine translation and multilingual language models (**litschko2021?**).

Recent advances in multilingual language models like mBERT (**devlin2019?**) and XLM-R (**conneau2020?**) have shown promise for cross-lingual applications, including for some low-resource languages. However, as (**lauscher2020?**) demonstrate, these models still show significant performance gaps for truly low-resource languages, particularly those that are typologically distant from high-resource languages.

In the scientific domain, (**nakov2018?**) explored cross-lingual biomedical information retrieval, highlighting both the potential and limitations of current approaches. Their work suggests that domain-specific knowledge and terminology present particular challenges for cross-lingual access to scientific content.

For COVID-19 specifically, (**guo2021?**) developed a multilingual search system for COVID-19 literature, but noted significant challenges for low-resource languages due to limited training data and linguistic resources. Their work underscores the need for better understanding of language representation in COVID-19 literature as a foundation for developing more inclusive information access systems.

0.2.5 2.5 Literature Gap

While previous research has examined language disparities in scientific publishing broadly and COVID-19 public health communications specifically, there remains a significant gap in understanding the linguistic diversity of comprehensive COVID-19 scientific literature collections like CORD-19. Additionally, few studies have specifically addressed representation gaps for low-resource languages, particularly African languages, in this context.

This research aims to address these gaps by providing a systematic analysis of language distribution in the CORD-19 dataset, with specific attention to low-resource languages. By examining not only language presence but also content characteristics, text complexity, and named entity distributions, this study will provide a more nuanced understanding of representation gaps that can inform future efforts to improve cross-lingual scientific information access.

0.3 3. Methodology

0.3.1 3.1 Research Design

This study will employ a mixed-methods approach combining quantitative content analysis with computational linguistic techniques to analyze the linguistic diversity of the CORD-19 dataset. The research design involves:

1. Automated language identification of documents in the dataset
2. Quantitative analysis of language distribution patterns
3. Content analysis of topic distribution across languages
4. Linguistic analysis of text complexity and named entity distribution
5. Comparative analysis to identify representation gaps for low-resource languages

This design allows for both broad quantitative assessment of language representation and deeper qualitative insights into specific gaps and barriers for low-resource languages.

0.3.2 3.2 Data Source

The primary data source for this study is the COVID-19 Open Research Dataset (CORD-19), a comprehensive collection of scholarly articles about COVID-19 and related coronaviruses. As of its latest release, CORD-19 contains over 1,000,000 scholarly articles, including over 400,000 with full text.

The dataset includes articles from various sources, including PubMed Central, bioRxiv, and medRxiv, as well as content from the WHO and commercial publishers. It provides both metadata (authors, publication date, journal, etc.) and full-text content where available, allowing for comprehensive language analysis.

0.3.3 3.3 Sampling Strategy

Given the size of the CORD-19 dataset and the computational resources required for language identification and analysis, this study will employ a stratified random sampling approach. The sampling strategy will include:

1. Initial random sampling of 10,000 documents to establish baseline language distribution
2. Stratified sampling based on publication date to ensure temporal representation across the pandemic
3. Targeted oversampling of potential non-English content based on metadata indicators (author affiliations, publication venue, etc.)
4. Full analysis of all identified non-English content to maximize insights on linguistic diversity

This approach balances computational feasibility with the need for comprehensive coverage, particularly for low-frequency languages.

0.3.4 3.4 Data Collection Procedures

Data collection will involve the following steps:

1. Acquisition of the latest version of the CORD-19 dataset
2. Extraction of relevant metadata (publication date, title, authors, abstract, full text if available)
3. Pre-processing of text content (cleaning, normalization)
4. Application of language identification algorithms to determine the primary language of each document
5. Collection of additional metrics (text length, publication venue, citation count if available)
6. Documentation of data processing decisions and limitations

0.3.5 3.5 Data Analysis Methods

0.3.5.1 3.5.1 Language Identification

Accurate language identification is crucial for this study. To maximize accuracy across both high and low-resource languages, multiple language identification approaches will be employed:

1. FastText language identification (**joulin2017?**) as the primary method due to its coverage of 176 languages
2. langdetect (based on Google's language detection) as a supplementary method
3. African Language Identification (AfriLID) tool (**adebara2022?**) for improved detection of African languages
4. Manual verification for a subset of documents to validate automated identification accuracy

For documents with multiple languages (e.g., abstract in multiple languages), the system will identify and record all languages present and their respective proportions.

0.3.5.2 3.5.2 Language Distribution Analysis

Quantitative analysis of language distribution will include:

1. Frequency counts and percentages of documents by primary language
2. Temporal analysis of language distribution across the pandemic timeline
3. Analysis of language distribution by document section (title, abstract, full text)
4. Correlation analysis between language and other document characteristics (publication venue, citation count, etc.)
5. Comparative analysis with language speaker populations to identify representational disparities

0.3.5.3 3.5.3 Content Type Analysis

To understand how content differs across languages, basic topic modeling and content classification will be performed:

1. Extraction of keywords and phrases using TF-IDF scores
2. Application of Latent Dirichlet Allocation (LDA) for topic modeling of English content

3. Classification of documents into broad categories (clinical, epidemiological, biological, etc.)
4. Cross-lingual comparison of topic distribution using available metadata and translated keywords

0.3.5.4 3.5.4 Text Complexity Analysis

Text complexity analysis will assess potential barriers to information accessibility:

1. Calculation of readability metrics (where applicable for supported languages)
2. Analysis of sentence length and syntactic complexity
3. Assessment of specialized terminology density
4. Comparison of complexity metrics across languages and document types

0.3.5.5 3.5.5 Named Entity Analysis

Named entity recognition will be used to identify terminology gaps:

1. Application of biomedical named entity recognition to English documents
2. Extraction of key medical entities (diseases, treatments, genes, etc.)
3. Analysis of entity distribution across available non-English content
4. Identification of terminology gaps for potential translation priorities

0.3.6 3.6 Ethical Considerations

While this study primarily analyzes publicly available dataset content rather than human subjects, several ethical considerations will be addressed:

1. Proper attribution of the CORD-19 dataset and compliance with its usage terms
2. Careful interpretation of findings to avoid reinforcing language hierarchies or stereotypes
3. Acknowledgment of limitations in language identification accuracy, particularly for low-resource languages
4. Transparent reporting of methodological decisions and their potential impact on results
5. Commitment to making findings accessible to diverse linguistic communities through translation of key results

0.3.7 3.7 Reliability and Validity

To ensure the reliability and validity of findings, the following measures will be implemented:

1. Multiple language identification methods with cross-validation to improve accuracy
2. Manual verification of a sample of language identification results
3. Transparent documentation of all data processing steps and analytical decisions
4. Acknowledgment of limitations and potential biases in the dataset and analysis methods
5. Validation of content analysis results through expert review where feasible

0.4 4. Expected Results

0.4.1 4.1 Anticipated Findings

Based on previous literature on language distribution in scientific publishing, the study anticipates finding:

1. Significant overrepresentation of English relative to global speaker populations, with potentially over 90% of content in English
2. Limited representation of major world languages like Chinese, Spanish, French, and Arabic
3. Minimal to non-existent representation of low-resource languages, particularly African languages
4. Potential temporal trends showing increased linguistic diversity later in the pandemic as information dissemination efforts expanded
5. Correlation between language availability and document characteristics such as publication venue prestige and citation count
6. Variations in content focus across different languages, with potential gaps in specialized topic areas for non-English content
7. Higher text complexity in non-English content, potentially reflecting translation of more complex scientific material

8. Significant terminology gaps for technical and biomedical terms in low-resource languages

0.4.2 4.2 Contribution to Knowledge

The expected results will contribute to knowledge in several ways:

1. Provide empirical evidence of language representation gaps in COVID-19 scientific literature
2. Quantify the extent of these gaps for specific language groups, particularly African languages
3. Identify patterns in content availability across languages that can inform translation priorities
4. Highlight specific barriers to information accessibility for speakers of low-resource languages
5. Create a methodological framework for analyzing linguistic diversity in scientific literature collections
6. Establish baseline measures that can inform future work on cross-lingual information access

0.5 5. Timeline

The proposed research will be completed within a 4-week timeframe with the following schedule:

0.5.1 Week 1: Data Preparation and Language Identification

- Acquire and preprocess the CORD-19 dataset
- Implement and validate language identification methods
- Create data processing pipeline for analysis
- Begin initial language distribution analysis

0.5.2 Week 2: Quantitative Analysis

- Complete language distribution analysis
- Conduct temporal and correlation analyses
- Begin content type classification
- Prepare preliminary findings on language representation

0.5.3 Week 3: Linguistic Analysis

- Conduct text complexity analysis
- Implement named entity recognition
- Analyze terminology distribution
- Begin comparative analysis of representation gaps

0.5.4 Week 4: Synthesis and Reporting

- Complete all analyses
- Synthesize findings across analytical components
- Prepare visualizations and tables
- Finalize research report and documentation

0.6 6. Limitations

This study has several limitations that should be acknowledged:

1. **Language Identification Accuracy**: Automated language identification may have lower accuracy for low-resource languages or short text segments. This limitation will be partially addressed through multiple identification methods and manual verification, but some misclassification may remain.

2. **Dataset Bias**: The CORD-19 dataset itself may have collection biases that affect language representation. For example, the methods used to collect articles may favor certain sources or languages.

3. **Content Analysis Depth**: Given the timeframe and resources, content analysis will be relatively high-level and may not capture nuanced differences in how topics are discussed across languages.

4. **Named Entity Recognition Limitations**: Biomedical named entity recognition tools are primarily developed for English and a few high-resource languages, limiting the depth of terminology gap analysis for low-resource languages.

5. **Temporal Coverage**: The analysis covers a specific time period defined by the CORD-19 dataset and may not reflect the most recent trends in linguistic diversity.

Despite these limitations, the study will provide valuable insights into linguistic diversity and representation gaps in COVID-19 scientific literature, with important implications for information access equity.

0.7 7. Conclusion

This research proposal outlines a systematic approach to analyzing the linguistic diversity of the CORD-19 dataset and identifying representation gaps for low-resource languages. By combining language identification techniques with content analysis and linguistic assessment, the study will provide empirical evidence of language-based disparities in access to COVID-19 scientific information.

The findings will have implications for global health information equity and will inform strategies for improving cross-lingual information access, particularly for speakers of low-resource languages. This research also provides a foundation for future work on developing better cross-lingual information retrieval systems and prioritizing translation efforts for scientific content.

Understanding the linguistic landscape of COVID-19 research is a crucial step toward ensuring that scientific knowledge is accessible to all language communities, regardless of the resources available for their languages. This study contributes to this goal by systematically documenting existing gaps and establishing priorities for addressing them.

1 References

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of ACL 2020.*