DECISION SCIENCES INSTITUTE

Mapping the Path to Translations: Empowering SIL to Reach New Frontiers in Language Translations

Vandana Brahmasa, Amy Fischbach, Parvathy Nair, Gitanjali Nambiar, Aryan Ritesh, Huilin Xu, Rohan Ajay, Matthew A. Lanham
Purdue University Daniels School of Business
vbrahmas@purdue.edu; fischba@purdue.edu; nair221@purdue.edu; nambiarg@purdue.edu; aritesh@purdue.edu; xu1997@purdue.edu; rajay@purdue.edu; lanhamm@purdue.edu

ABSTRACT

We developed a mapping tool for SIL International, a global leader in language translation, to identify unmet translation needs and guide market expansion efforts. The motivation for this research is that despite global connectivity, many low-resource languages lack Bible translations; SIL seeks to address this gap using AI and data science. We design and build a mapping tool with Python and HTML which identifies underserved areas, integrates LLMs to help identify barriers, and offers insights applicable to both SIL and broader global outreach initiatives.

KEYWORDS: Language Translation, Low-Resource Languages, Market Expansion, Large Language Models, Data-Driven Mapping

INTRODUCTION

In today's rapidly evolving world, information flows smoothly, aided by advanced technologies and global communication networks. Despite these advancements, certain communities remain underserved due to the absence of resources in their native languages. This disparity becomes especially noticeable regarding vital documents such as legal guidelines, educational materials, and religious or faith-based texts. According to a survey done by Wycliff, only 10% of the languages have a fully translated Bible. Religion is an integral part of many individuals' day-to-day lives, highlighted in many of the decisions they make and comfort they feel. For Christians, connecting to the Word of God is an important aspect of their faith. A study published by Jonas S. Thinane highlights how important Bible translation is to the mission of the church and their community. When entire communities lack access to foundational content, they face barriers in literacy, cultural preservation, and participation in broader social dialogues. One significant challenge lies in prioritizing which languages should receive translated resources first, given the constraints of both financial and human capital. A major nonprofit organization, SIL International, seeks to maximize the impact of its translation efforts. SIL focuses on bridging the divide by bringing key texts, particularly foundational religious materials, to communities lacking translations.

The business problem in our study centers on identifying which language communities should be prioritized to ensure the most beneficial and timely return on investment. Many factors influence this prioritization. Geographic challenges, for instance, can make travel and on-site work prohibitively expensive. Similarly, languages that are closely related may allow for quicker translation adaptations, while those that are more distinct may require a full translation project from the ground up. Additionally, demographic indicators, such as the size of a language community or its socioeconomic conditions, play a role in determining both feasibility and long-term benefits. By assessing these interrelated variables, planners can arrive at more strategic

decisions about where to focus immediate resources. Many companies, including SIL, seek a systematic, data-driven approach to guide their strategic planning. Decision-makers must also be able to visualize relationships between communities, calculate travel costs, evaluate linguistic distance, and weigh social or cultural impact. Graph-based analytics tools, such as the isochrone graph, offer a potential solution, as they can integrate disparate data sources into a single computational framework. In such a system, individual language communities become nodes in a network, while edges between them can represent everything from geographic proximity to linguistic similarity. This model allows for the calculation of "distances" in multiple senses—whether physical, linguistic, or socio-cultural—and thus enables sophisticated analyses.

Our primary research question is: How can a flexible, graph-based planning tool integrate these factors to help an organization allocate its limited resources more effectively? While the resulting strategic planning tool focuses on our industry collaborator, SIL, the design and technologies can be extended to other markets – retail, manufacturing, etc. All firms need the ability to make better decisions today but also adapt to future changes in travel routes, population shifts, and technological breakthroughs. The creation of this tool is not only to serve the specific needs of SIL, but also to be a framework for modifying it to address various other business needs as well. Isochrones maps are maps that depict the area accessible from a point within a certain time threshold, and are being utilized by companies in a variety of industries to address their geographic concerns. To address this business need, this study proposes the development and deployment of related geo-maps. Our tool aggregates information on travel time, cost estimates, linguistic distances, and demographic metrics in a way that is both dynamic and scalable. The goal is to present findings through intuitive data visualizations and dashboards that allow non-technical stakeholders to grasp complex networks and support data-driven decisions.

LITERATURE REVIEW

Bible Translation: History and Benefits

Bible translation (BT) into local languages and dialects has been a significant preoccupation of missionary work and has played a significant role in revitalizing distinct linguistic traditions. BT has seen various phases of evolution, from being an endeavor in word-for-word equivalence to an exercise in capturing dynamic equivalence with diverse languages (Naude, 2005). The Third Generation of Bible translations has expanded focus to incorporating oral traditions in its translations, which are seen to be more palatable to some communities. According to Naude, subsequent scripture translation should be motivated by balancing the need for amenability to chanting with literary quality and explainers on cultural context. Scripture translation has been accompanied by myriad socio-economic transformations in the communities it has been carried out, markedly in the spheres of education, literacy and health outcomes.

Barriers to Translation

Language translations are mediated by several factors, the most fundamental among them being the impact of geography on translation practices. Geographical contexts form the backdrop of development of different languages and linguistic traditions, which in turn pose a challenge in interpretability of translated works. Effective translation is thus an endeavor in balancing the source content's message with the target audience's interpretations, often through personalized metaphors rooted in the target audiences' cultural context (Ke, 2019).

Beyond the linguistic challenges in translation, there are practical socio-economic, cultural and political factors that mediate access to a region to allow translation efforts to flourish. The Skopos theory (Du, 2012) of language translation suggests that translation is a goal-driven activity primarily guided by the translator's intent.

Spatial Visualization Techniques

Isochrones can offer an intuitive way to visualize the regions that have available translations and the degrees to which translations are available. Isochrones are lines that connect points of equal travel time (Desai, 2008). In other words, it shows us the regions that can be reached in the same amount of time. Isochrones allow us to visualize the accessibility of a region and can be enriched with contextual inputs like mode of transport, speed, infrastructure available to glean insights into the accessibility of a location.

In addition to isochrone mapping, this paper utilizes Voronoi diagrams for visualizations of regions with and without scripture translations. Voronoi diagrams geometrically divide a plane into different regions based on their distance from 'seeds'- given points on the map. A Voronoi diagram divides the plane into several cells that enclose a region closest to each seed (DeLorenzo, 2021). For purposes of this study, the regions which have scripture translations available may be considered 'seeds' of the Voronoi cell. Each Voronoi cell would thus represent the wider region in which the language is spoken, a geospatial data point that can guide not only scripture translation efforts, but other region-determined endeavors like community building, business operations and expansion, tourism and aid work.

DATA

Bible translation data used in this data came from publicly available data at Progress.Bible (https://progress.bible/data/). As shown in Table 1, this dataset contained information on various regions and languages, along with the status and date of Bible translations available for each.

The ProgressBible dataset captures key details about Bible translation efforts across different languages and regions. It includes the country name (Country) and the language spoken (Language Name), along with the size of the language population (Population Group) and the stage of life for the language (Language Vitality). It also specifies which pieces of the Bible have been translated (Scripture), when it was published (Year Scripture Published), and whether there is an active Bible translation underway (Active Translation).

This data served as the foundation for our research and allowed us to assess the extent of Bible translation efforts across different linguistic and geographical contexts. By analyzing these attributes, we aimed to identify patterns in recent translation activity, gaps in coverage, and potential areas where additional translation efforts might be needed.

Companies will often use additional datasets to improve transparency from other angles. We integrated additional information from external sources, including geographic coordinates (latitude and longitude) for each region. For accurate location data, we utilized Glottolog (https://glottolog.org/), a comprehensive linguistic database that provides detailed information on the world's languages, including their geographic distribution. Glottolog is maintained by the Max Planck Institute for Evolutionary Anthropology and follows rigorous scholarly standards, ensuring the accuracy and reliability of its linguistic and geographic data. This expansion facilitated geo-based mapping, allowing for more precise visualization and assessment of translation distribution.

To estimate distance and travel time between locations, we incorporated the OSRM API (https://project-osrm.org/). Adding the Ease of Doing Business Index published by the World Bank Group added another intelligence layer (World Bank, 2020). This index evaluates the performance of different countries relative to each other on metrics like ease of starting a

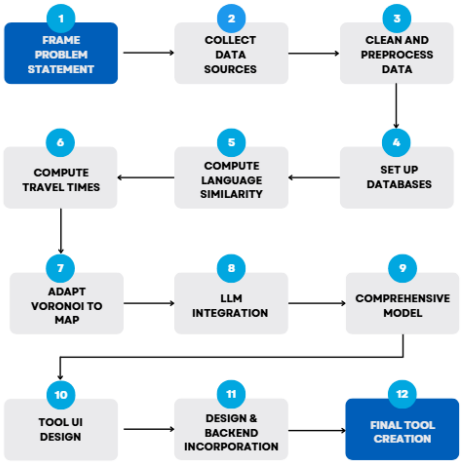
business, enforcing contracts, access to electricity, etc. By combining linguistic and geographical data, our tool has more potential to provide comprehensive insights about the accessibility and reach of Bible translations worldwide.

METHODOLOGY

To create a visualization tool for mapping locations with and without scripture translation, we first conducted market research to locate various sources of open-source data that could be utilized for this mapping tool. A rigorous review of literature was conducted to identify barriers to entry for Bible translations, as well as strategic and logistical challenges in accessing different linguistic regions. After narrowing down our sources of data, we proceeded to clean and preprocess the data to make it usable for analysis.

For language distance, we computed the similarity of languages with each other and crystallized it to a language similarity score. Our approach leverages lang2vec, which is built on the URIEL database, a large-scale linguistic resource that is structured compendium of information on language typology (https://aclanthology.org/E17-2002/). The Lang2vec library provides language embeddings based on these features allowing computational comparisons across languages. Travel times were computed, i.e. the time it takes to travel to a certain region using the OpenSky (https://openskynetwork.github.io/opensky-api/rest.html) and OSRM APIs. With these contours of our mapping tool in place, we created the necessary composite database to be utilized by the tool. Figure 1 below shows our step-by-step methodological workflow.

Figure 1: Methodological workflow



GEO-MAPPING TOOL DEVELOPMENT

To enhance the visualization of linguistic regions serviced by translation efforts, we employed a Voronoi-based geospatial mapping approach. This method partitions geographical areas based on linguistic distribution, effectively delineating regions where translations are available and those that remain underserved. Given the global scope of translation expansion, we opted for a world map visualization, ensuring a comprehensive representation of translation accessibility across diverse regions. Additionally, we prioritized user accessibility, incorporating an interactive feature that allows users to input a language of interest, which serves as the reference point for analysis.

The core functionality of the tool revolves around identifying the nearest hub language, defined as the geographically closest language to the input language that possesses an existing translation. To facilitate intuitive interpretation, we employed a color-coded classification system, highlighting gaps in translation coverage, as we aim to reveal regions with high linguistic accessibility versus those still in need of translation efforts.
We also sought to enhance the analytical capabilities of the tool by integrating external data sources to provide a more comprehensive assessment of translation accessibility. Apart from external APIs, we also used church location coordinates, sourced from OpenStreetMap (https://www.openstreetmap.org) (OSM), as religious institutions often play a pivotal role in translation dissemination and linguistic accessibility. The inclusion of this dataset provides a more holistic perspective, allowing users to assess translation accessibility within the context of community infrastructure.

To improve the interpretability of global translation accessibility, we implemented dynamic clustering (with the level of zoom) in our tool. When the map is displayed, languages that are in proximity are clustered together and assigned a representative color to have an immediate visual summary of translation density, allowing users to quickly identify high-priority regions for translation expansion:

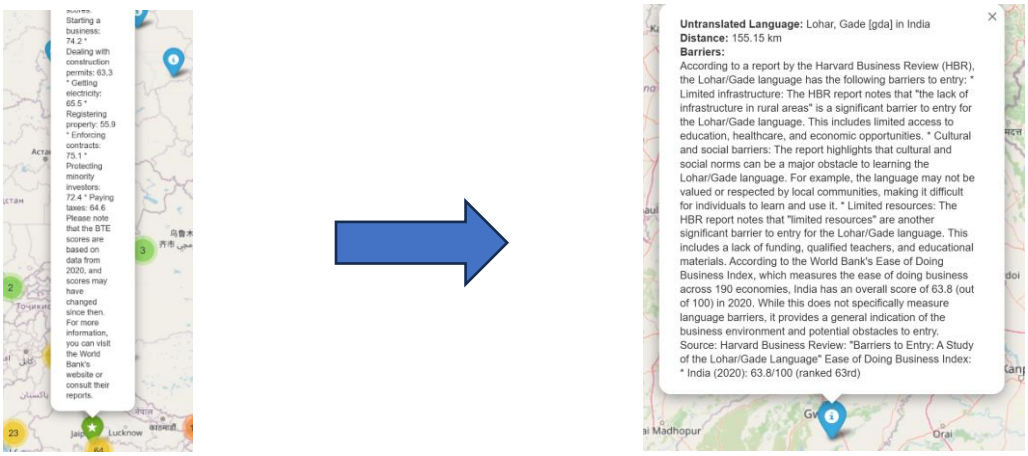
- Red clusters represent regions where most languages remain untranslated.
- Yellow clusters indicate regions with a moderate balance between translated and untranslated languages.
- Green clusters highlight areas where most languages have existing translations.

To ensure an accurate representation of linguistic distributions, we overlaid a Voronoi diagram onto the global map, using the hub languages as input points. Since Voronoi diagrams naturally extend infinitely across the mapped space, we constrained the visualization to landmasses by integrating a shapefile-based clipping mechanism. To achieve this, we incorporated shapefiles from Natural Earth (https://www.naturalearthdata.com/downloads/10m-physical-vectors/), a public domain geographic dataset. This adjustment ensures that Voronoi regions align with real-world geographical boundaries, preventing distortions caused by oceans.

Leveraging additional data sources, we incorporated several critical analytical components to provide a more comprehensive evaluation of translation accessibility. Based on the user's input language, the tool provides: Distance to the nearest translation hub, estimated travel time to the nearest hub, identification of the closest church location, distance to the closest church, estimated travel time to the closest church.

Building out the LLM portion of our tool went through multiple iterations to provide a format that worked well with our final goal. In using a localized LLM, the prompt needed to be extremely specific to show the barriers in a usable fashion. At first, the paragraph showed to be too long to be readable, along with not correctly identifying the language input by the user. Adjusting the prompt to focus on only the top three points helped shorten the output. To address the issue with identifying the language, the for-loop in Python was adjusted to first find the language name from the excel datasheet and then use that name in the chat prompt. Addressing the validity of the LLM output was also a concern, so the prompt was anchored with a reputable news source, like the Harvard Business Review, to be sure that it was reporting on barriers provided by a trusted source. This is demonstrated in Figure 2.

Figure 2: Example of information provided



Data Analysis

Using the tool, we observed some interesting trends that could help SIL. To narrow down a starting point, we can see overall which regions contain the most unreached people groups, which is South Asia as shown in Figure 3.

Figure 3: Underreached markets

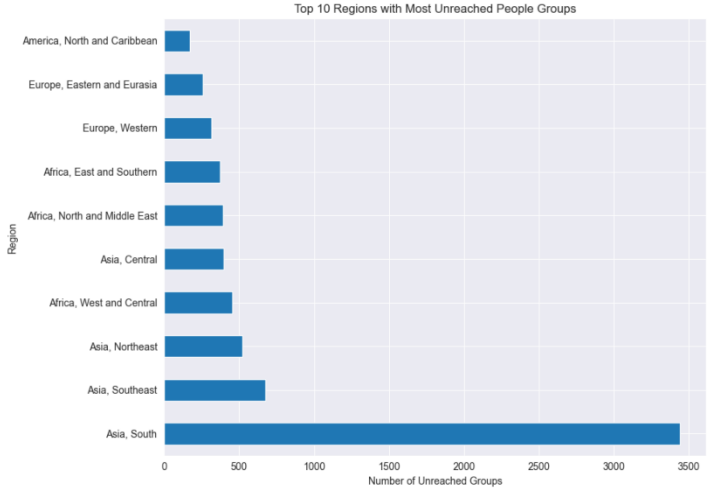
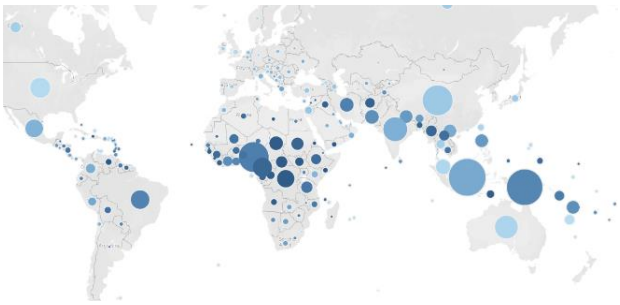
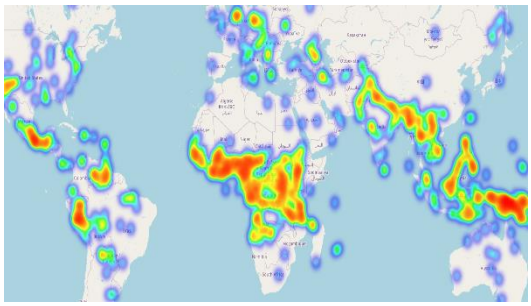


Figure 4 shows how the map on the left can be used to see more specific areas with untranslated languages, shown by the size of the circles, combined with the coloring of the circles representing the Ease of Doing Business score of that country. The lower the ranking, the darker the color is. The map on the right in Figure 4 shows where active translations are happening.

Figure 4: Mapping functionality highlights key opportunity areas

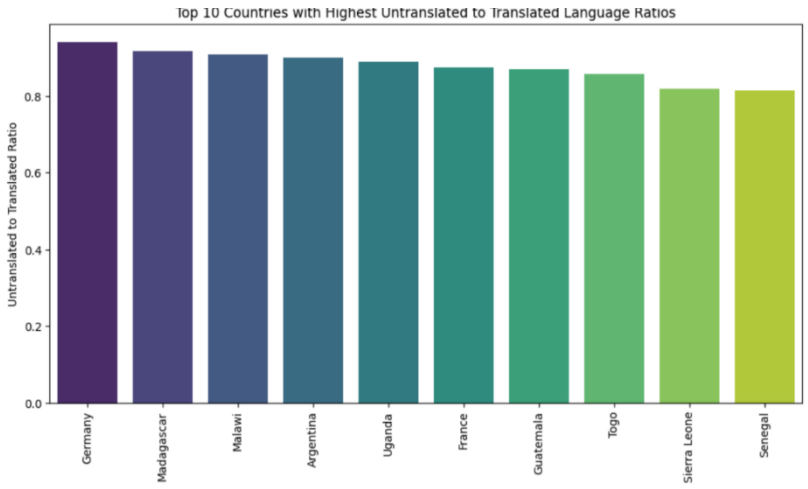




These maps highlight areas that show high numbers of untranslated languages, like the middle of Africa and Southeast Asia and show where efforts are already being made. One can note that while some are still untranslated, some progress is being made to start there. This can help turn attention towards relatively untapped areas, such as China and Brazil. Using the coloring from the EODB score, SIL can see that China has a lower score vs Brazil, and thus potentially making it an easier country to focus on.

We observed from Figure 5 that Germany has the highest untranslated to translated ratio, meaning that while it does have a well-known global language, many of the languages spoken in the region remain without a Bible translation. This would make Germany another good hub to use as an anchor point, as it's a country with a higher EODB score and who's primary language is already translated.

Figure 5: Top 10 countries with highest untranslated to translated language ratios



The LLM coding for the barriers to entry also showed us that the top three barriers to entry for languages were linguistic complexity, governmental or religious restrictions, and resources as shown in Figure 6.

Figure 6: Top three identified barriers

This is shown by the fact that many languages with smaller people groups that speak it often have less available written resources that can be utilized for translation purposes. In addition, these regions often do not have Christianity as their primary religion, which can lead to an uphill battle with entering the country to do translations, especially if the government is heavily tied to a different religion. With these challenges, it often means more resources would be necessary to address these concerns, which is difficult in an industry that already has a limited number of resources to utilize. What this tells us is understanding the trends in the tool will help SIL and others address these three big concerns to make a strategic next step.

REFERENCES

Cronin, M. (2003). Translation and globalization. https://doi.org/10.4324/9780203102893.ch36

de Lorenzo, S. (2021). Generalized Voronoi Diagrams: Theory and Related Applications. University of Salzburg

Desai, Kiran. (17 October 2008). Isochrones: Analysis of Local Geographic Markets (PDF). Mayer Brown. Retrieved 2018-05-31

Dovey, K., Woodcock, I., & Pike, L. (2017). Isochrone Mapping of Urban Transport: Car-dependency, Mode-choice and Design Research. *Planning Practice & Research*, 32(4), 402–416. https://doi.org/10.1080/02697459.2017.1329487

Du, X. (2012). A brief introduction of Skopos theory. *Theory and Practice in Language Studies*, 2(10), 2189-2193. https://doi.org/10.4304/tpls.2.10.2189-2193

Ke, X. (2019). Instruction of Tourism English from the Perspective of Translation Geography

Lehmann, U., Dieleman, M., & Martineau, T. (2008). Staffing remote rural areas in middle- and low-income countries: A literature review of attraction and retention. *BMC Health Services Research*, 8, 19 - 19. https://doi.org/10.1186/1472-6963-8-19

Mahendradhata, Y., Kalbarczyk, A. (2021). Prioritizing knowledge translation in low- and middle-income countries to support pandemic response and preparedness. *Health Res Policy Sys*, 19, 5. https://doi.org/10.1186/s12961-020-00670-1

Naudé, J. A. (2005). On the Threshold of the Next Generation of Bible Translations: Issues and Trends. *Meta*, 50(4). https://doi.org/10.7202/019851ar

Thinane, J. S. (2024). Translating missio Dei: Indispensable Bible translation in God's mission. *Verbum et Ecclesia*, 45(1), Article a2841. https://doi.org/10.4102/ve.v45i1.2841

World Bank. (2020). Ease of Doing Business Methodology. World Bank Group

Wycliffe Bible Translators UK. (n.d.). Record-breaking year. Retrieved March 21, 2025, from https://wycliffe.org.uk/story/record-breaking-year