

---

# DATA WHITEPAPER

JULY 2023

**PREPARED BY:**  
ETEN INNOVATION LAB

---

# Introduction

The ETEN (Every Tribe Every Nation) Innovation Lab is dedicated to accelerating Bible translation toward the 2033 All Access Goals (AAG) by making use of advanced technology. Two factors influencing the usefulness of advanced technology to this end are the degree to which it can be integrated into tools that are helpful for translation teams and the existence of enough quality data in AAG languages to build the technology. This data guide describes the language data needs for the Innovation Lab. It introduces the various problems that could prevent us from achieving our mission, some solutions to those problems, realistic goals, the relevant stakeholders who will accomplish these goals, and a summary of the research and development which must be done.

## Challenges

The following section lists the problems we must overcome in order to get the language data we need for our work. These are several existential threats as well as factors contributing to the challenge of gathering the data necessary for meeting the ETEN AAGs using advanced technology:

1. **Stakeholder rejection:** The government, leaders, community, or translation team could say “no” to us. Without local support, we will not be able to collect data or implement any accelerating solutions we develop.
2. **Stakeholder ambivalence:** The communities might stop providing data. Without the correct incentives for speakers to continue providing data, we will not be able to make any progress.
3. **Wrong data:** Certain sets of gathered data might be useless for ETEN’s purposes. If the data we collect is not of the right type or of a sufficient amount, we will not be able to make any progress.
4. **Existence of diverse data:** AAG languages will all have different starting points from a data perspective. This makes a one-size-fits-all approach unrealistic. This means that our solutions must take these different starting points into account.
  - Some will have no data.
  - Some will have some non-Scripture data.
  - Some will have some Scripture data that can be used.
  - Some will have a whole New Testament that can be used.
  - Some will have high linguistic similarities to languages that already have high digital language support and/or Scripture that can be used.
  - Some will have data in a variety of formats and modalities (audio, text, video, images formatted in Word documents, USFM, text files, PDFs, audio files, etc.).
5. **Degrees of standardization:** Many of the AAG languages are non-standardized with respect to spelling. Many will not even have standardized orthography. It is traditionally a

very time-consuming process to establish an orthography and get the community to agree on standardization.

6. **The dialect continuum:** Many of the AAG languages will have multiple dialects with a continuum of mutual comprehensibility. This means that data collection will need to involve certain steps toward standardization and decision-making about which dialect to select for Biblical materials.
7. **Oral-first/oral-only languages:** Many AAG languages are oral-first/oral-only languages. This means that these languages cannot take advantage of many tools in traditional Bible translation.

## Solutions

1. **Stakeholder rejection:** We use trusted partners and create ways for the government, leaders, and the community to see the local benefit. To this point, Daniel Wilson and his team at XRI and ILAD (International Literacy and Development) have multiple years of experience in providing the types of solutions that mitigate this threat. They are working in multiple countries that are hostile to faith-based organizations. They are also working with many universities, Google, and UNESCO, and are even receiving funding from secular organizations to provide their language-related services—services which provide local benefits to the communities. Their reputation, experience, and broad goals enable them to be an important source of data for this project.

Other partners which can deliver on the data needs of the Innovations Lab are translation service providers such as [Xelex](#). Sourcing, preprocessing, and labeling data is their expertise. They are well suited to provide data for many of the AAG languages and incentivize/manage rapid collection efforts.

2. **Stakeholder ambivalence:** Our partners use financial and social incentives to motivate speakers to continue to provide data. Partners, such as Xelex and XRI, provide financial incentives for data providers for their methods. XRI employs gamification as well to keep users interested in continuing to contribute data.
3. **Wrong data:** The data plan will be strategic and based on empirical research on the best data for the output that ETEN expects. The teams will be agile and will work together in multiple simultaneous and inter-connected tracks. The scope and scale of this project means that, practically, there must be multiple tracks working in parallel to one another. We must have teams doing research, collecting data, testing hypotheses, testing integrations, testing quality, and developing additional technology simultaneously. This will ensure that we know precisely what data we need at all times for all languages.
4. **Existence of diverse data:** The different starting points for the different languages can be addressed by conducting a language data assessment at the beginning of the process. This assessment will take inventory of all useful data, including data from related languages and their [level of digital language support](#). Following this assessment, we can leverage multimodal techniques to make use of the existing data available in a language.

5. **Degrees of standardization:** This issue can be addressed in a few ways. For written languages, the community can take advantage of in-app voting mechanisms that could make this process more efficient. The community will make decisions about the standardization of spelling. The community may prefer oral in those cases where the standardization process is less desirable and/or impractical.
6. **The dialect continuum:** It is possible to use technology-assisted adaptation for different dialects if this is the best course of action for a given cluster of languages/dialects. There have been attempts to take a completed translation in one language and use technology to adapt that translation into a very closely related language/dialect (e.g. [Scripture Forge](#) and [FlexTrans](#)). This is possible since the degree of variation between languages/dialects is limited and measurable. The data we are collecting can illuminate precisely how the languages and dialects differ and aid in the development of tools for creating good adaptations via [dialect identification](#) in data and [transfer learning](#) for similar languages.
7. **Oral-first/oral-only languages:** Oral Bible translation (OBT) methods and tools such as [Render](#) have significantly helped these languages receive Scripture. However, there are bottlenecks in these processes, which can be accelerated through advanced technologies as we collect audio data.

The task of accessing this amount of data in so many languages is massive. This means that it will be very strategic to run several different approaches concurrently. A multi-pronged, multi-provider approach with fast cycle times and multiple iterations is a best-case scenario for addressing the data needs for this many languages and mitigating the challenges listed above. The best approaches to sourcing language data will rise to the top and be deployed quickly into new contexts.

## Goals

Given the diversity of the types of languages and their varying levels of standardization and digital language support, there are several useful benchmarks for text-based translation projects that could be established:<sup>1</sup>

1. **Corpus readiness:** This means that a dataset that is sufficient, preprocessed, and in domain exists for a language. This dataset must have a large enough parallel corpus of in-domain data that is ready to train a model. Corpus readiness will be an iterative process which will require multiple cycles with different types of data and parameters.
2. **Dictionary readiness:** This means that the semantic concepts that are present in the desired portions of Scripture have been collected and preprocessed.
3. **Integration readiness:** This means that the language model and other materials have been prepared to integrate with select tools. It also means that our partners are ready to integrate *what* we have developed into their tools.

---

<sup>1</sup> These benchmarks speak to text-based translations. Different benchmarks must be developed for oral-based translations.

With respect to text-based projects, deliverables for the 2a and 2b sections of the Innovation Lab can be set based on these benchmarks. Languages will move through different stages of completion with respect to these benchmarks. A strategic effort will be needed to contact and prepare partners for any technology solutions we provide for any languages. Partners include communities and translation teams on the ground as well as technology providers such as Scripture Forge.

The table below illustrates how we could plan, measure, and execute by benchmark (rows) and by stage (columns). We can accurately measure progress by tracking how many languages have identified partners who are ready to help with data collection for the corpus readiness and dictionary readiness benchmarks. Likewise, once the partners have been identified and data collection has begun, the language project can move to the “in progress” stage. Once the languages have reached the “completed” stage for both the corpus readiness and dictionary readiness benchmarks, they can begin the “integration process” stage (green row).

We expect that we can efficiently approach questions about timelines and cost by projecting increases in the number of languages in each of these stages for the given benchmarks.

	Partners ready	In progress	Completed
Corpus readiness	5	3	2
Dictionary readiness	7	4	6
Integration readiness	6	4	2

These benchmarks are helpful for goal setting.

**Partners**

This project depends on the contributions of multiple partners simultaneously addressing different aspects of the overall project. The following work is already in motion by our partners:

- Daniel Whitenack (SIL) has produced important [Hot Spots](#) research that demonstrates, to those working in the data collection initiative, which languages will likely yield the earliest and best results.
- He has also conducted experiments on the [Golden Path](#), demonstrating that the order in which the translation is done can have a beneficial effect on the quality of language models.
- Both Randall Tan ([Clear](#)) and Daniel Whitenack with AQuA have produced important quality assurance tools.
- Daniel Wilson ([XRI](#)) and his team have [deployed multiple types of apps](#) in multiple locations to collect language data which is used to build the type of technology the Innovations Lab needs.

- The TRE initiative, led by Tim Jore, will benefit from this data initiative since it will help with their goals to get resource materials translated into the various resource levels.
- The Innovations Lab has already secured partnerships with Scripture Forge, Paratext, and Faith Comes by Hearing and has started discussing integrations. We will be able to test initial integrations simultaneously with the other activities.
- Daniel Whitenack (SIL) and Randall Tan (Clear) have started working with [Xelex](#) to test their capacity for sourcing, preprocessing, and labeling text data via landing pages, paid advertising, and financial incentives.

## Research and Development

The research section provides a description of the types of data that are necessary for this project as well as how we approach data quality. The development section provides more detail about the tools we will use for accomplishing our goals.

### Research

#### Text Data

Text data consists of any sentences or words which have been harvested for the purpose of building dictionaries or language models that are useful for integrating with Bible translation tools. Tools such as Scripture Forge or Paratext already have remarkable capabilities with sufficient data. We expect that these tools can have additional capabilities which will provide translation suggestions and even generate batch translations if sufficient data has been supplied.

For text data to be valuable for our purposes, it must (1) be able to bootstrap a translation suggestion or MT (machine translation) system to produce first drafts or suggestions that speed up drafting by a significant amount and (2) be able to bootstrap quality estimation techniques that speed up quality estimation by a significant amount. We believe that we will need to rapidly update our NLP (natural language processing) methods in the coming years, but at this stage we believe that text data which will be valuable for our purposes must have these three qualities:

1. Sufficient semantic coverage: This means that the semantic concepts which exist in the portions of Scripture in view must be represented in the dataset. If the goal is to translate Luke 5 into one of the languages on the AAG list, the semantic concepts present in Luke 5 must be present in the dataset.
2. Sufficient grammatical coverage: This means that the grammatical and syntactic elements which exist in the portions of Scripture in view must have been learned by the language model. This means that there is sufficient data in the datasets which contain the grammatical and syntactic peculiarities of that portion of Scripture.
3. Sufficient domain adherence: This means that the data collected is domain specific to the domain of the Bible and uses appropriate spiritual, religious, etc. terminology, rather than

purely general domain language or language from other domains such as health, business, etc.

## Audio Data

Audio data will be valuable for our purposes for different reasons than text data. We anticipate that audio data will be most useful for oral-first/oral-preferred languages which will likely be receiving Scripture in audio form. This especially includes OBT work. We anticipate audio data assisting in at least two ways:

1. Key term spotting for OBT consultants. With enough audio data, we can train a model to identify and flag a key term while working with a tool such as [Render](#) so that a consultant can ask specific questions about why a certain key term appears or does not appear in sections of Scripture where it is or is not expected. Ideally the consultant and translation team could request spotting for additional words or phrases to be integrated into quality assessment tools such as AQuA.
2. Automatic speech recognition systems can be created for these languages which will allow a text-based rendering of the oral Scriptures, enabling consultants to use traditional tools on oral languages.

Below is a list of other datasets which will be valuable for specific translation projects:

1. Word alignment data.
2. Cluster specific language dataset with language/dialect identifications.
3. Datasets of key biblical terms.
4. Updated comprehension questions. We have pretty good sets from AQuA, unfoldingWord, and Transclerator, but these sets still need filtering and translation into other languages.
5. Languages of wider communication (LWCs) of AAG languages and existing Bibles.
6. Pronunciation dictionaries (aligning text with corresponding phones).
7. Audio files with aligned, transcribed text.
8. Annotated sign language videos.
9. Translation memory: A database that stores sentences, paragraphs, or segments of text that have been previously translated.
10. Aligned Greek and Hebrew datasets with LWCs.

## Data Quality

We know that in order to build machine translation and quality estimation systems, quality language data must exist for both source and target languages. The performance of any machine translation and quality estimation system is [greatly dependent on the amount of training data available in that language](#).

We know that data from related languages [can improve model performance](#). The [Language Hot Spots research](#) done by Daniel Whitenack (SIL) and his team has demonstrated that many of the AAG languages are related to languages with high digital support. This means that once we have

some data in the target language, we believe we can leverage data from the related languages to train machine translation systems more quickly. This helps us identify which AAG languages to start with.

We know that data quality is also an important factor in the performance of a machine translation system. Data quality is influenced by the [in-domain nature of the training data](#). When using crowdsourcing techniques to source language data, it is also critical to avoid “[translationese](#)”, preferring natural language data instead. These two factors, among others, will greatly influence the type of data we collect so we can ensure high-quality data.

In addition to the baseline experiments already planned to be run on the e-bible corpus within PAB-NLP, we are guided by previous studies which provide details about the amount of data, type of data, and model performance for low-resource languages such as the following:

- [Emakhuwa-Portuguese](#)
- [English-Yoruba](#)
- [MT for related Southern African Languages](#)
- [Bambara-French](#)
- [Empirical Analysis of Parallel Corpora and In-Depth Analysis Using LIWC](#)

We also need to answer the following questions:

- What is the ideal approach to collecting this data?
- What level of incentives are needed to obtain the minimum target dataset?
- How do we leverage data from clusters of related languages for ETEN’s purposes?

We have started to address these questions in the proposal *Collecting and Analyzing Multilingual Language Data*, which was funded by Biblica. This research project will provide empirical evidence for how well language models perform when trained on different types of data gathered in different ways.

## Development

### **Apps/Web Apps**

To achieve corpus readiness and dictionary readiness benchmarks, a mobile app-based approach is strategic and has been successful in multiple locations for collecting useful data. For example, see [this effort](#) which used a WhatsApp bot to collect 26,240 parallel sentences in a low-resource language in a few months. When combined with other incentives, such as micropayments and gamification designs, we believe the pace and amount of data can be increased.

The 2b section of the Innovation Lab has already built a team and certain components which will likely be useful in the development of crowdsourcing tools for language data. We plan to build upon what they have developed where it complements the specific tech stack we need. We also plan to coordinate our teams’ efforts, utilizing their expertise in crowdsourcing.

Our partner, XRI, will deploy apps in partnership with the Innovations Lab. The goal is to produce a small number of app templates that are optimized for the type of language data we are collecting (text vs. audio) which can be reused in each location, after localizing the interface and prompts. The reuse of these app templates will greatly reduce the time it will take to deploy the technology into new contexts.

## **Database**

Certainly, there are important decisions yet to be made around the structure and storage of the data. It is likely that the Innovation Lab will store the data and models for languages (in the TRE) which have completed both corpus readiness and dictionary readiness benchmarks. That said, there are many questions to explore in order to identify what approaches would be optimal in this area. Our priority in structuring and storing data is to radically broaden involvement and increase acceleration.

## **Integration**

We will work closely with the teams focused on building the language models and integrating with Bible translation tools. We will create a workflow that is optimized for downstream integration with the NLP layer, Scripture Forge, AQuA, Paratext, and other tools which will benefit from our data. By collaborating early and often with these other teams, we will be able to iterate quickly.

## **Conclusion**

The field of natural language processing is evolving rapidly. This document reflects our approach as of Q1 2023 to source the data we need to accelerate Bible translation through advanced technologies. We are committed to keeping up with developments in the field and pivoting as necessary in order to meet the goals ETEN has given us.