Association of Petroleum Geologists. India
Registered under Societies Registration Act,No. 21 of 1860

GEO India 2022

An effective way of using LaTex for Typesetting Indian Regional Languages

Santosh Kumar Sahu
Email: sahu_santosh@ongc.co.in, Oil and Natural Gas Corporation Limited

Abstract

Word processing applications offer limited utility while rendering scientific and mathematical documents. Though they provide many features related to writing equations, tables, images, references, and bibliography management, the typographic quality is often compromised. In addition, writing the Indic language, especially writing ligature words, often changes while using multiple systems to prepare the document. The characters are changed by switching the application or operating system. Therefore, in this study, we will discuss how to write the Indic languages more effective and efficient way that can easier to prepare high-quality scientific documents. Also, it is easier to transfer or use in multi-working environments. LaTex typesetting system is used to prepare the Indic languages such as Hindi, Odia, Telugu, Malayalam, and Bengali. Fontspec and Polyglossia LaTex packages are used to render the regional languages. It is observed that these two packages are very effective to prepare the Indic languages using a single Font namely Nirmala UI that supports more than 10 Indian languages. We have used ligature words in equations, tables, figures, and references using these six Indic languages more easily and effectively to prepare the technical and scientific report.

Keywords*:* Latex, Indic Language, Hindi typesetting, Odia typesetting, Telugu typesetting, Malayalam typesetting, Bengali typesetting, Fontspec, Polyglossia

Introduction

Before the development of the Unicode System, many proprietary software was used to write the Indic languages. Due to the different methods used to develop the software, there is no compatibility in sharing the project used in one writing system with another. As a result, users are forced to use the specific application for word processing in Indian regional languages. But nowadays we are using the Unicode encoding method which supports 144,697 characters to represent many languages and symbols that are used for word processing or writing systems (Unicode 14, 2006). Still, we face problems by typesetting and sharing a document that contains a huge amount of ligature words, symbols, and equations in scientific and technical reports (Alex, 2006). As a result, we will use the LaTex typesetting system to write the Indic languages in a better and more effective way.

The word processor allows users to see the document in real time while writing. It is so effective for preparing simple and small documents (Datta, 2017). For scientific writing, a lot of effort is required to write complex math symbols, equations, images, and tables, and manage bibliography contents. Figure 1 shows the effort versus the complexity of a document for technical and scientific writing. Initially, there is some effort required to learn the basic syntax and semantics of LaTex, and after that one can easily be prepared a high-quality document. Therefore, in this study, we will discuss how can write Indic languages more efficiently and effectively using LaTex. How LaTex can be used to prepare the Indian regional languages and for the font required in his experiment has been discussed by (Kumar, 2016). There are 11 types of fonts used in his experiment to write different Indic scripts.

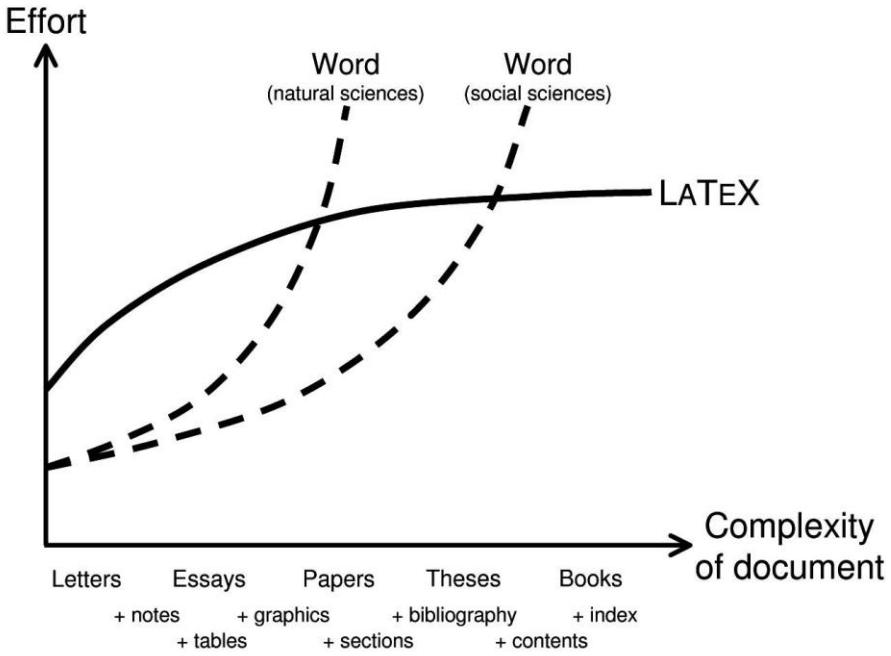

Figure 1: Efforts required with respect to the complexity of the document (Efforts vs Complexity, 2022)

Scientific Publishing in India

India has become the world's third-largest publisher of technical and scientific articles and published approximately 1.35 lakh papers (Publication Ranking, 2022). The number of publication is drastically increased as per the statistics of the National Science Foundation(NSF), US. To prepare high-quality and impressive articles, reports, and dissertations, the LaTex typesetting system is used. Therefore, this study helps Indian authors to prepare good standard reports, or manuscripts in regional languages using LaTex.

The advantages/disadvantages of LaTex over word processing applications are as follows:

A. Advantages

- Professionally crafted templates are available that make the document more impressive.
- During writing Equations, a huge number of symbols are available that can easily integrate with the document.
- Managing bibliographic content and reference becomes easier and more convenient.
- It emphasizes writing the content, and LaTex cares about visualization.
- It encourages the writer to prepare a well-structured document by using predefined templates.
- Many typesetting engines are available that support a wide variety of packages to perform specific tasks.
- It is open source and anyone can modify, and customize the MiKTeX compilers and packages as per requirement.
- Support Windows, Mac, and Linux Operating systems as well as web servers.

B. Disadvantages

- Learning syntax and Symantec of LaTex is time-consuming.
- Very hard to prepare unstructured documents without a template.
- The error handling is very poor. Even a single mistake, a lot of errors will generate in the error log.

LaTex and Indian Languages

Despite the discouraging situation in Latex usage in India, some reputed universities motivated their students to write their reports, papers, and thesis using LaTex. Templates are created by many publishers to learn and prepare the document more easily. A lot of websites are available that provide troubleshooting during scripting, some provide services like online table maker, figure generator, and many more. In this study, we have simulated their procedure to prepare a document in Hindi with Devanagari Font. we have observed that in the case of the mixed font (language i.e., Hindi and English) the characters of the English font are displayed with a square symbol. The actual character is not visible. This has also happened in the case of symbols and special characters like commas, semi-colons, and full stops.

Omega package for typesetting for use of the Malayalam language is used by (Alex, 2006). The UTF-8 encoding scheme is used and two fonts are used namely Keli and Rachana in his experiment. Only the Malayalam Indic language is considered in his experiment. Hence, in this study, we will discuss how to typeset multiple Indic languages using a single font namely Nirmala UI developed by Microsoft (Nirmal UI, 2022).

The remaining part of the paper is organized as follows. Section 2 discussed the methodology used in this study, Section 3 presents the results of the study and Section 4 concludes and future scope of this experiment.

Methods

The objective of this study is to install new TrueType fonts that support Indic languages in the LaTex typesetting system. The detailed system requirements are discussed in Table 1. MikTex installed on Windows 10 with TexStudio as Tex Editor. After those two fonts mentioned in Table 1 are installed. The preamble of the document used in this study is shown in Figure 2. Six Indic scripts are declared with two fonts. The Devanagari package is used to simulate the work by earlier authors. The issue with using this package is that each time only one language option will be set. Commonly, the numbers and dates are written in English. As a result, to typeset, the Hindi fonts and English fonts were separated by curli braces. To avoid such type syntax, in this study, an alternative approach is used to typeset the Indic Languages.

Table 1 System Requirements

Operating System	Windows 10
LaTex Compiler	MikTex
Typesetting Engine	XeLatex
Integrated Writing Environment	TexStudio
Font	Nirmala UI
Packages	Fontspec and Polyglossia

The language of the document is classified into two types default language and another language. After that, the languages and their corresponding font are declared. In this experiment only a single font i.e., Nirmala UI is used. The body part of the document is shown in Figure 3. In the whole-body section of the document, whenever any Indic language is required to use, simply start with the selection language command with the language name as shown in Figure 3. For English, there is no requirement of declared the selection language macro.

```
\usepackage{polyglossia}
\usepackage{devanagari}
\usepackage{fontspec}


\setdefaultlanguage{english}
\setotherlanguages{hindi, odia, tamil, bengali, telugu, malayalam}
\newfontfamily\devanagarifont[Script=Devanagari]{Nirmala UI}
\newfontfamily\odiafont[Script=Odia]{Nirmala UI}
\newfontfamily\tamilfont[Script=Tamil]{Nirmala UI}
\newfontfamily\bengalifont[Script=Bengali]{Nirmala UI}
\newfontfamily\telugufont[Script=Telugu]{Nirmala UI}
\newfontfamily\malayalamfont[Script=Malayalam]{Nirmala UI}
```

Figure 2: Preamble section of the study

The output of the LaTex script is shown in Figure 4. All the Indic scripts are displayed followed by an English sentence. The intention of writing an English sentence along with the regional language is to show that using this method we can generate Indic scripts in English.

```
\par \textbf{English}
\par Knowledge is the supreme goal

\medskip

\textbf{Hindi}

\par \selectlanguage{hindi}
ज्ञान सर्वोच्च लक्ष्य है | Knowledge is the supreme goal
\medskip

\textbf{Odia}
\par \selectlanguage{odia}
ଜ୍ଞାନ ହେଉଛି ସର୍ବୋଚ୍ଚ ଲକ୍ଷ୍ୟ | Knowledge is the supreme goal

\medskip
\textbf{Bengali}
\par \selectlanguage{bengali}
জ্ঞানই সর্বোচ্চ লক্ষ্য। Knowledge is the supreme goal

\medskip

\textbf{Tamil}
\par \selectlanguage{tamil}
அறிவு மிகச் சிறந்த இலக்கு| Knowledge is the supreme goal
\medskip
```

Figure 3: Body part of the document

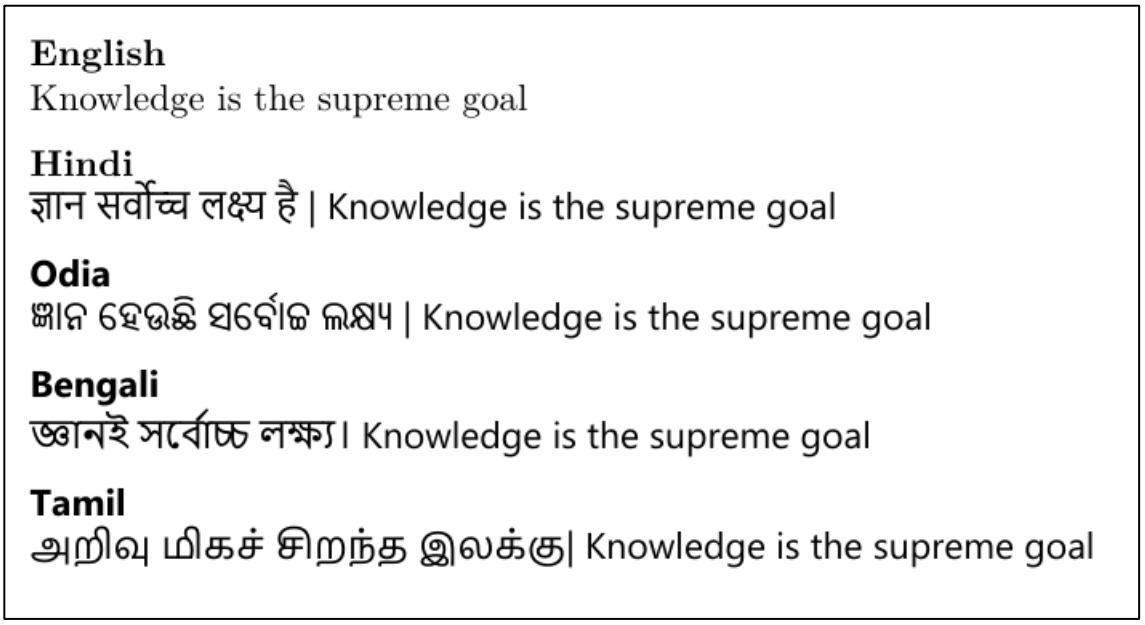

Figure 4: Indic language output as per our study

Results and Discussion

In the previous works, while writing multiple languages such as English and Hindi in a document, the English characters are shown using the symbolic format. To avoid that one can split the sentences and declare the English characters separately. To avoid such situations in this study, the recently developed Nirmala UI font used that support more than 10 languages. As per Figure 4, each regional language is followed by a sentence in English. No need split the sentences in our experiment. One can insert a single paragraph or multiple paragraphs at a time inside the definition of the particular language code shown in Figure 3. Therefore, using this approach, one can use English words and sentences within any regional language.

Conclusion

In this study, the LaTex typesetting system is used to prepare Indic languages. Two latex packages namely Fontspec and Polyglossia are used. Nirmala UI font developed by Microsoft that supports more than 10 Indian languages. Therefore, this font alone is used in this experiment to prepare documents in multiple languages. Earlier techniques discussed in the paper are based on a single font single language concept. It was difficult to follow the syntax for individual languages which is overcome by using the Nirmala UI font. In our future work, other packages and fonts that support other Indic languages will simulate using the LaTex typesetting system.

Acknowledgment

This work would not have been possible without the support of our Head of the Institute, GEOPIC, ONGC. We would like to express my very great appreciation to Manoj Ranjan, GM(Geol.), and Sankhadip Bhattacharya, SG for their valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated. We are also immensely grateful to all GEOPICians for their support and encouragement to complete the study.

References

Unicode 14.0.0., 2006, https://www.unicode.org/versions/Unicode14.0.0/ (accessed Jul. 10, 2022).

Alex, A. J., 2006, Typesetting Malayalam using Ω. [Online]. Available, URL: http://sarovar.org/projects/malayalam

Datta, D., 2017, LaTeX in 24 hours: a practical guide for scientific writing. Springer., DOI: 10.1007/978-3-319-47831-9.

Kumar R., 2016, Latex and Indian Languages Creating a Cloud-based Library Catalogue using Google Fusion Tables View project Implementation of Virtual Learning Management System at Tumkur University View project LaTeX and Indian Languages, in Knowledge Utsav - National Conference, pp. 1–4. DOI: 10.13140/RG.2.1.2978.9840.

Effort vs complexity, 2022, the preparation of scientific and technical writing. https://softwareengineering.stackexchange.com/questions/47402/what-is-the-best-toolkit-for-writing-long-technical-texts (accessed Jul. 10, 2022).

Publication Ranking, 2022, India is the world's third largest producer of scientific articles: Report - The Economic Times. https://economictimes.indiatimes.com/news/science/india-is-worlds-third-largest-producer-of-scientific-articles-report/articleshow/72868640.cms?from=mdr (accessed Jul. 14, 2022).

Nirmala UI font family, 2022 - Typography | Microsoft Docs. https://docs.microsoft.com/en-us/typography/font-list/nirmala-ui (accessed Jul. 10, 2022)