

# **Architecting Socio-linguistic AI-models Facilitating Accessible and Engageable Digital Communication for Semi-literates**

**THESIS**

*Submitted in partial fulfilment of the requirements for the degree of*

**DOCTOR OF PHILOSOPHY**

*by*

**PRAWAAL**  
(2018PHXF0701P)

*Under the Supervision of*

**PROFESSOR NAVNEET GOYAL** (BITS Pilani)

**PROFESSOR POONAM GOYAL** (BITS Pilani)

**DR. VINAY M. R.** (Infosys Bangalore)



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE,  
PILANI**

**November 2024**



*Empowering the under-represented and marginalized,  
with technology aids,  
is not just a matter of digital inclusion,  
access and engagement.*

*It's about unlocking the limitless human potential,  
within every individual,  
fostering opportunities,  
and building a more inclusive future for everyone.*



## *Declaration of Authorship*

I, Prawaal, declare that this thesis titled, 'Architecting Socio-linguistic AI-models Facilitating Accessible and Engageable Digital Communication for Semi-literates' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

**Date**

\_\_\_\_ / \_\_\_\_ / \_\_\_\_

---

**PRAWAAL**

2018PHXF0701P

BITS Pilani, Pilani Campus



**BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI  
(RAJASTHAN)**

*Certificate*

This is to certify that the thesis entitled "Architecting Socio-linguistic AI-models Facilitating Accessible and Engageable Digital Communication for Semi-literates", submitted by Prawaal ID.No 2018PHXF0701P for the award of Ph.D. degree of the Institute, embodies original work done by him under our supervision.

\_\_\_\_\_  
(Signature of the Supervisor)

**PROFESSOR NAVNEET GOYAL**  
Senior Professor,  
Department of CSIS,  
BITS Pilani, Pilani Campus

Date: \_\_\_\_\_

*Dr. Vinay M. R.*

\_\_\_\_\_  
(Signature of the Co-Supervisor)

**PROFESSOR POONAM GOYAL**  
Professor,  
Department of CSIS,  
BITS Pilani, Pilani Campus

Date: \_\_\_\_\_

\_\_\_\_\_  
(Signature of the Co-Supervisor)

**DR. VINAY M. R.**  
Senior Data Scientist,  
Infosys,  
Bangalore, Karnataka, India

Date: \_\_\_\_\_





*Dedicated to my parents, wife and my curious son.*



## *Acknowledgements*

*“Gratitude is the fairest blossom which springs from the soul.”*

---

Henry Ward Beecher

First and foremost, I express my gratitude to the divine forces for imbuing me with unwavering spirit and resilience throughout this journey. This has been a source of strength, keeping my motivation high even during the most challenging times. Additionally, the creation of this thesis is attributed to the invaluable support extended by numerous individuals, and I extend my heartfelt appreciation to each one of them.

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Navneet Goyal, for his invaluable guidance, persistent support, and constant encouragement throughout my doctoral journey. His expertise, wisdom, and depth of knowledge in the field of Artificial Intelligence (AI) and Natural Language Processing (NLP) have been instrumental in shaping my research and academic growth. I am truly grateful for the opportunity to work under his mentorship.

I am equally thankful to my co-supervisor, Professor Poonam Goyal for her invaluable mentorship, constructive feedback, and dedication to fostering a nurturing academic environment. Her meticulous review of our papers, insightful comments and thorough feedback were instrumental in the successful publication of our work. Without her dedicated review and thoughtful contributions, these publications would not have come to life. I am deeply appreciative of her encouragement, guidance, and belief in my abilities, which have instilled confidence and resilience in me throughout this doctoral journey.

I am deeply thankful to both the supervisors for their understanding of my professional commitments at my workplace. They provided invaluable support and guidance, allowing me to effectively balance my work responsibilities with my PhD studies. Their flexibility in scheduling meetings, accommodating video conferences, and facilitating remote communication enabled me to seamlessly navigate both realms and make progress in my academic pursuits while fulfilling my professional obligations.

I am also thankful to Dr. Vinay MR, my co-supervisor at my workplace, for his local support, guidance, and collaboration on the day to day basis. He has facilitated and greatly enriched my research work and helped me during the entire journey of research.

I would like to express my heartfelt gratitude to the Departmental Research Committee (DRC) convener, members of the DRC, the Dissertation Advisory Committee (DAC), and the esteemed leadership of BITS Pilani for their unwavering support and invaluable guidance throughout my PhD journey.

Last but not the least, I am indebted to my family for their unwavering love, encouragement, and sacrifices throughout this journey. Their unwavering support and understanding have been my source of strength and motivation.

**Prawaal**

**ID. 2018PHXF0701P**

## Abstract

In an increasingly digitized world, the access to technology has become integral to daily life, from accessing information to communication and collaboration with others. People in developing countries particularly those without tertiary education, or other accessibility challenges face hurdles in using digital platforms for communication and collaboration. The linguistic diversity of this section of population across the world, makes design of near-universal digital enablement methodology a challenging task. The inability to use digital technology for these individuals can exacerbate existing inequalities and limit opportunities for social, economic, and educational advancement. These individuals form the epicenter of our work, and refer them as "*semi-literates*" in our entire work.

Building accessible and engageable digital assistants for semi-literates will need an ecosystem of tools and techniques thereby building an all-inclusive digital world for everyone. Architecting AI-models for this population needs deep analysis of socio-linguistic challenges, along with data-driven analytics and inferencing to understand the issues and expectations of these groups of people. It is expected that not all semi-literates would face same challenges, and hence it is very important to categorize this group into smaller subgroups based on their needs and wants and design technology driven models for each subgroup separately.

In our work, we conduct an extensive fieldwork collecting text messages from multiple demographics (urban and rural) exchanged between semi-literates. Through the detailed analysis of these messages, we categorize the population of interest into three distinct groups based on their behavioral patterns using a variety of Artificial Intelligence (AI) and Machine Learning (ML) algorithms. Categorization of semi-literates into smaller groups enables a more nuanced understanding (challenges, preferences and needs), and the need for separate models for each to enhance digital accessibility and inclusivity for each group.

The first group of semi-literates, referred as "*Digitally-Novice*" in our work comprises of individuals who possess basic digital literacy however require assistance in the reduction of language complexity on digital platforms. The second group, termed as "*Digitally-Niche*", consists of individuals who are proficient in local languages and express a preference for digital content in their native languages. Finally, the third group designated as "*Digitally-Neglected*", comprises of individuals with minimal exposure to orthographic text, necessitating the most assistance in navigating digital content.

*Digitally-Novice* individuals encounter challenges in comprehending information available on the internet and other digital platforms, primarily due to the semantic and syntactic complexity inherent in language constructs. The majority of content found on digital platforms is tailored towards proficient users, employing enriched vocabulary not appropriate for semi-literate individuals. To address this issue, we propose a reductive paraphrasing framework specifically tailored for digitally-novice users, leveraging principles from Neural Machine Translation (NMT) and late averaging transformer architecture. Our methodology incorporates Zero-shot Learning (ZSL) utilizing a multi-pivot approach to execute our experiment effectively. We employ creative use of data, by building the simplified vocabulary from the animation movies targeted for younger audiences. Through empirical testing, we ascertain a noteworthy enhancement of around 25% in the ease of comprehension. This enhancement facilitates Digitally-Novice individuals in accessing and navigating digital content with increased ease and proficiency. We further elaborate on this work in Chapter 3.

The irony for *Digitally-Niche* semi-literates lies in the inability to access digital information due to language barriers. While this group is proficient in their native language, however there is no digital content (or insufficient content) for the languages of these groups on digital ecosystem. Architecting AI-models for marginalised languages face a dual, first is the lack of foundational systems such as Optical Character Recognition (OCR) and font designs and the second is the lack of high-quality large-scale datasets to train deep learning models effectively. In order to address these challenges comprehensively, we propose two distinct solutions. Firstly, we present a versatile OCR methodology tailored for ultra-low resource languages, leveraging AI methods like contrastive learning, GAN-augmentation, and auto glyph feature extraction and bring digital on-boarding for these language and scripts (further illustrated in Chapter 4). Secondly, we introduce a scalable and algorithmic dataset-generation framework designed to cater to the needs of low resource languages, utilizing cross-lingual embeddings and Content-Based Image Retrieval (CBIR) techniques. Our novel framework extracts bilingual parallel corpora from newspaper articles, facilitating the creation of high-quality datasets essential

for training deep learning models (further illustrated in Chapter 5).

The last category we refer as *Digitally-Neglected* semi-literates, are the ones who need the most assistance for digital enablement. This group faces significant challenges in reading traditional orthographic scripts and have bare-minimum academic literacy. This group of semi-literates face utmost difficulty with the use of digital devices, due to their non-exposure to written forms of communication. These individuals hardly are able to read anything beyond numbers and small words. To mitigate these difficulties, we propose a novel approach involving the development of a multimodal ideographic script enriched with spatial and directional attributes. This metalanguage is designed to transcend linguistic barriers and minimize the risk of misinterpretation. We conducted empirical evaluations along with qualitative feedback from members of this group. We observe an accuracy exceeding 80% in the comprehension of semantic elements across diverse linguistic backgrounds (further illustrated in Chapter 6).

To conclude, this thesis attempts at the digital literacy enablement, particularly for semi-literate individuals of varying degrees as described above. Through the integration of advanced technologies such as Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP), alongside state-of-the-art solutions including Generative AI, this research has provided innovative approaches to address the multifaceted barriers encountered by such populations. Leveraging the AI-models proposed in our work is not just a step towards unlocking the human potential for these groups, but will also foster a deeper understanding of the complex interplay between technology and socio-linguistic patterns. The findings presented herein underscore the potential of AI-driven interventions to empower semi-literate individuals with diverse backgrounds, thereby promoting digital inclusivity.





## *Abbreviations*

AAC	Augmentative and Alternative Communication
AI	Artificial Intelligence
ARIA	Average Reading Index Improvement
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CASI	Center for the Advanced Study of India
CBIR	Content Based Image Retrieval
CCA	Connected Component Analysis
CCL	Connected Component Labeling
CER	Character Error Rate
COT	Chain Of Thoughts
CW	Complex Words
CWI	Complex Words Identification
DA	Data Augmentation
E2E	End To End
EHX	Enhanced Horizontal Projection
ELMo	Embeddings from Language Models
ETSI	European Telecommunications Standard Institute
FIRE	Forum for Information Retrieval and Evaluation
GAN	Generative Adversarial Networks
GDPR	General Data Protection Regulation
GPT	Generative pretrained Transformer
GFRS	Glyph Feature Recommendation System
Gen-AI	Generative Artificial Intelligence
HP	Himachal Pradesh
HCI	Human Computer Interaction
HX	Horizontal Projection
ICTCS	Information and Communication Technology for Competitive Strategies
IDD	Intellectual or Developmental Disabilities

IOT	Internet Of Things
LAS	LaBSE Alignment Score
LLM	Large Language Model
LRL	Low Resource Language
LRS	Low Resource Scripts
MIA	Multiple Index Approach
ML	Machine Learning
MSM	Multimodal Semantic Metalanguage
MT	Machine Translation
MTT	Meaning-Text Theory
NLP	Natural Language Processing
NMT	Neural Machine Translation
NP	Noun Project
NSM	Natural Semantic Metalanguage
OCR	Optical Character Recognition
OPUS	Open Parallel Corpus
OOV	Out Of Vocabulary
PII	Personally Identifiable Information
POS	Parts Of Speech
PRImA	Pattern Recognition and Image Analysis Research Lab
RAG	Retrieval Augmented Generation
ROI	Region Of Interest
SARI	Systematic Assessment of Reading Information
SC	Semantic Class
SIFT	Scale-Invariant Feature Transform
SLAS	Sentence Length Alignment Score
SLT	Semi-literate Texting
SM	Semantic Molecule
SOTA	State Of The Art
SSE	Sum of Squared Errors
ST	Semantic Template
STE	Simplified Technical English
STS	Semantic Textual Similarity
SURF	Speeded-Up Robust Feature
SV	Semantic Variable
TOT	Tree Of Thoughts
TTP	Text To Picture
UDL	Universal Design for Learning

UI	User Interface
UX	User Experience
W2V	Word To Vec
WER	Word Error Rate
ZSL	Zero Shot Learning



## *Symbols and Notations*

$\forall, \in, \notin$	The symbol $\forall$ is read as "for-all" or "for-each", and $\in$ is read as "belongs-to" or "in". $\notin$ is just opposite of $\in$ .
$\Sigma, \Pi$	The capital sigma symbol $\Sigma$ (pronounced "sum") is used to denote summation in mathematical notation. $\Pi$ (pronounced "product") denote multiplication of a group of entities. Both of these symbols are used with limits to demote the range of the group to be summed/multiplied.
$\emptyset, \exists, \forall, \cup, \cap, \setminus$	These are the symbols used in set theory and logical operations.
$=, \neq, \approx, \leq, \geq, \ll, \gg, \subset, \subseteq, \supset, \supseteq$	These symbols represents relationships like equality, inequality, and subset relationships.
$+, -, \times, \div, \cdot, \pm, \mp, \cdot, *, \div$	These are mathematical operators and are used for addition, subtraction, multiplication, division, and other operations.
$\alpha, \beta, \gamma, \delta, \varepsilon, \theta, \lambda, \mu, \phi, \pi, \tau, \omega$	These are other Greek letters commonly used in mathematics and science often used to represent variables, constants, and other mathematical entities and are described individually when used within equations.
$\rightarrow, \leftarrow, \Rightarrow, \Leftarrow, \leftrightarrow, \Leftrightarrow$	Symbols for arrows and implications are used in logical expressions and diagrams.

## List of Figures

1.1	A cumulative and recursive model of successive kinds of access to digital technologies. . . . .	2
1.2	Adoption of digital platforms geographically . . . . .	3
1.3	Adoption of mobile phones geographically (in millions) . . . . .	4
1.4	The reach of digital platforms across various countries. . . . .	5
1.5	Internet growth in India, in the last 15 years. . . . .	6
1.6	Linguistic distribution of web content. . . . .	7
1.7	Most widely spoken languages worldwide [11]. . . . .	8
1.8	Native languages in India with more than 30 million users. . . . .	9
1.9	Familiarity with English across (a) Demographics (b) Social discrimination (c) Education and (b) Economic prosperity . . . . .	10
1.10	The long-tail of linguistic distribution on the internet. . . . .	15
2.1	Heatmap for various parameters. . . . .	29
2.2	Analysis of conversations (a) Wordcloud - Rural (b) Wordcloud - Urban (c) Sentiment - Urban and (d) Sentiment - Rural . . . . .	32
2.3	Analysis of corpus on metadata and text message characteristics. . . . .	34
3.1	Schematic representation of relationship between three languages as outlined in the linguistic theory of NSM [19] . . . . .	46
3.2	Sentence alignment methodology across subtitle files of two languages.[77] . . . . .	50
3.3	High level flow diagram of Reductive Semantic Paraphraser . . . . .	52
3.4	Lexical Simplification procedure workflow . . . . .	53
3.5	NMT using Late Averaging and 3 languages pivot model . . . . .	55
4.1	High level design for Versatile unsupervised OCR methodology for Low-resource scripts Through Auto Glyph feature Extraction (VOLTAGE) . . . . .	69
4.2	Partition of characters into small groups based on glyph features as recommended by our design. . . . .	72

4.3	Synthetic data generated for Takri - (Left Image) applying GAN algorithm and (Right Image) applying Image transformations . . . . .	73
4.4	Use of supervised contrastive learning for symbol classification . . . . .	74
4.5	Use of foundational model and transfer learning using annotated Takri dataset . . . . .	77
4.6	Character error rates (CER) for machine printed test scenario leveraging CNN-LSTM models and compare the results with VOLTAGE models . . . . .	78
4.7	Character error rates (CER) for handwritten test scenario leveraging CNN-LSTM models and compare the results with VOLTAGE models . . . . .	79
5.1	Article mapping using images as pivots. . . . .	86
5.2	Proposed data augmentation architecture. While Data Article Extractors (Data & Article) work on the two languages independently, the Mappers (Article & Sentence) use the languages in conjunction. . . . .	87
5.3	Annotation nomenclature for various components in the article . . . . .	88
5.4	Extrinsic evaluation of dataset using Machine translation task . . . . .	96
6.1	Hierarchical structure of our ontology. . . . .	110
6.2	High level process on experimental design. . . . .	110
6.3	Final illustrative message (continued from Table 6.3) . . . . .	114
6.4	Ideographic volume following logarithmic growth with exploration of dataset. (x-axis is corpus size, and y-axis is ideographic count) . . . . .	115
6.5	Results for ideographic effectiveness using MIA . . . . .	118
8.1	Survey form for data collection - Page 1 . . . . .	128
8.2	Survey form for data collection - Page 2 . . . . .	129

## List of Tables

1.1	Paraphrase typology for various types . . . . .	13
1.2	Language Resource Distribution with number of languages and speakers. Category 0,1, 2 and 3 are referred to as Low Resource Languages (LRL) . . . . .	16
2.1	Survey data details - variables and their description . . . . .	31
2.2	Impact of gender, demography, age and profession on length of text messages (in characters). . . . .	33
2.3	Statistical Differences Between Clusters: T-Test P-Values . . . . .	35
3.1	Training dataset . . . . .	52
3.2	Illustrative examples of complex and converted paraphrased sentences from our work . . . . .	56
3.3	SARI scores for various state of the art models . . . . .	57
3.4	Impact of paraphrasing on readability indexes. . . . .	58
3.5	Comparing single Vs multi-pivot. ARII (Average Reading Index Improvement) is the average of all readability scores, to measure the simplicity of paraphrased sentences. . . . .	58
4.1	Unsupervised clustering accuracy for various zones for various k-means combinations. . . . .	71
4.2	Empirical study for VOLTAGE on Takri on Machine Printed (MP) and Hand Written (HW) samples. . . . .	75
4.3	Evaluation across other scripts. For Gujarati we experimented with two scenarios, (a) Gujarati LRL- Like low resource language and (b) Gujarati HRL- like high resource language . . . . .	76
5.1	Semantic Textual Similarity (STS) for different sentence lengths. LaBSE Alignment Score (LAS), Sentence Length Alignment Score (SLAS) and Lexical Overlap Alignment Score (F-Score) . . . . .	91
5.2	STS for different article sizes (in terms of number of sentences within the article) . . . . .	93



5.3	Spearman correlation between LAS and SLAS for multiple sizes of articles . . . . .	94
5.4	Corpus sentence count and sentence length at multiple headers . . . .	94
5.5	Cumulative Semantic Textual Similarity (STS) in final corpus . . . . .	95
5.6	Evaluation for translation task on Punjabi-Hindi combination . . . . .	95
6.1	Examples from our dataset, following our mathematical model . . . .	108
6.2	Semantic clustering approaches and findings. Human evaluation is normalised on 0-1 scale with 1 being best score. (W2V - Word2Vec, FT- FastText) . . . . .	108
6.3	Step by step illustrative message conversion. . . . .	113
6.4	Empirical Evaluation (on 0-1 scale) on A: Meteor, B: S-BERT, C: MPNet and D: mini-LM. . . . .	116
6.5	Empirical evaluation using multiple prompt engineering techniques. . . . .	117
6.6	Empirical Evaluation (on 0-10 scale) on A: Expressiveness, B: User Experience, C: Intention to Reuse and D: Interest. . . . .	118

# Table of Contents

<b>Declaration of Authorship</b>	<b>v</b>
<b>Certificate</b>	<b>vii</b>
<b>Acknowledgements</b>	<b>xi</b>
<b>Abstract</b>	<b>xiii</b>
<b>Abbreviations</b>	<b>xvii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	4
1.2 Problem Discovery and Design Principles . . . . .	7
1.3 A Comprehensive Literature Review for Enhancing Accessibility and Engagement in Lower Education Cohorts . . . . .	11
1.3.1 Text Simplification: Enhancing Comprehensibility along with preserving Semantic Integrity . . . . .	11
1.3.2 State-of-the-Art Techniques for Low Resource Languages . . . . .	15
1.3.3 Innovations in Visual Communication Methods for Enhanc- ing Accessibility in Low-Education Populations . . . . .	18
1.4 Research Objectives, Scope, and Boundaries . . . . .	20
1.5 Major Contributions . . . . .	22
1.6 Thesis Organization . . . . .	22
<b>Part 1 - The Groundwork</b>	<b>25</b>
<b>2 A Versatile Dataset for Socio Linguistic Assessment of Semi Literate     Urban and Rural Populations in India</b>	<b>27</b>
2.1 Semi-literate Texting (SLT): A dataset for the semi-literates, by the semi-literates . . . . .	27
2.2 SLT Dataset: Experimental Design, Materials and Methods . . . . .	29
2.3 Modeling of Semi-Literate Personas: Analysis and Insights from SLT	33
2.4 Persona Analysis: Defining Strategic Objectives for our work . . . . .	38

2.5	Ethics Statement . . . . .	39
<b>Part 2 - Enabling "Digitally-Novice" Semi Literates using Semantic Simplification</b>		<b>41</b>
<b>3</b>	<b>Multi-Pivot Sequence-to-Sequence Transformer with Late Fusion Architecture for Reductive Paraphrasing</b>	<b>43</b>
3.1	Introduction . . . . .	43
3.1.1	Defining Research Objectives: AI-Models for Enhanced Semantic Cohesion in Sentence Simplification . . . . .	44
3.2	Assessing State-of-the-Art Research Trends and Identifying Gaps in Reductive Paraphrasing Methods . . . . .	45
3.2.1	Review of SOTA Computational and Linguistic Methods . . . . .	45
3.3	Critical Analysis: Unaddressed Issues and Opportunities in Reductive Semantics . . . . .	48
3.4	Novel Zero-shot Reductive Paraphrasing Methodology using Multi-Pivot late fusion Transformer Model . . . . .	49
3.4.1	Semantic Ontology-Based Extraction and Construction of a Large-Scale Training Dataset . . . . .	49
3.4.2	Design of Multi-Pivot Zero-Shot Reductive Semantic Paraphraser . . . . .	52
3.5	Evaluation of proposed AI Model Performance on Effective Sentence Simplification and Preserving Semantic Integrity . . . . .	55
3.6	Conclusion and Limitations . . . . .	59
<b>Part 3 - Enabling "Digitally-Niche" Semi Literates through Automatic Data Generation and Augmentation</b>		<b>61</b>
<b>4</b>	<b>Versatile Low Resource OCR Methodology using Contrastive Learning, GAN Augmentation, and Auto Glyph Feature Extraction</b>	<b>63</b>
4.1	Introduction . . . . .	63
4.1.1	Defining Research Objectives: Towards Establishing Robust OCR AI-Models with Limited-to-None Digital Training Corpus . . . . .	64
4.2	State-of-the-Art in OCR: Literature Review and Gaps for Low-Resource Script Digitization . . . . .	65
4.3	Analyzing Deficiencies and Emerging Challenges in Low-Resource OCR Models . . . . .	66

4.4	VOLTAGE: Proposing a Novel and Versatile OCR Methodology for ultra low-resource scripts . . . . .	67
4.4.1	Strategic Selection of the Endangered Scripts for Experimental Focus . . . . .	67
4.4.2	VOLTAGE: Pipeline Design for ultra low resource scripts . . . . .	68
4.5	Analytical Insights and Discourse on VOLTAGE’s OCR Results for Endangered Scripts . . . . .	75
4.6	Conclusion and Limitations . . . . .	78
<b>5</b>	<b>Dataset Generation Framework using Cross-lingual Embeddings and Content Based Image Retrieval</b>	<b>81</b>
5.1	Introduction . . . . .	81
5.1.1	Defining Research Objectives: Multimodal AI for Automated Low-Resource Bilingual Parallel Dataset Creation . . . . .	82
5.2	Survey of existing AI-Powered Multimodal Strategies and Identifying Gaps for Generating Bilingual Parallel Datasets in Low-Resource settings . . . . .	83
5.3	Critical Assessment of Limitations and Challenges in Multimodal AI Approaches . . . . .	85
5.4	Proposed AI-Powered methodology for Low resource Bilingual Parallel Corpus Extraction . . . . .	85
5.4.1	The Design Intuition: Image-Pivoted Article Mapping and Advanced Sentence Embeddings . . . . .	85
5.4.2	Manifesting the Design Philosophy: Experimental Setup and Implementation Details . . . . .	86
5.5	Evaluation of proposed AI Model Performance on Multiple Language pairs . . . . .	90
5.5.1	The choice of Metrics . . . . .	90
5.5.2	Emperical Evaluation . . . . .	92
5.5.3	Case Study . . . . .	95
5.6	Conclusion and Limitations . . . . .	96
	<b>Part 4 - Enabling "Digitally-Neglected" Semi Literates Through Gen-AI</b>	<b>99</b>
<b>6</b>	<b>Generative AI powered Multimodal Semantographic Metalanguage</b>	<b>101</b>
6.1	Introduction . . . . .	101
6.1.1	Articulating Objectives: Multimodal Generative AI Metalanguage for Users with Limited Orthographic Script Familiarity . . . . .	102

6.2	Survey of current Ideographic Methods of Communication on Digital Platforms . . . . .	103
6.3	Uncovering Challenges and Opportunities in Multimodal Communication Methods . . . . .	104
6.4	MagicHood: Introducing Generative AI Powered Methodology for Hierarchical Ideographic Communication on Digital Platforms . . .	105
6.4.1	Strategic Integration of Ideography and Text in Multimodal Communication . . . . .	105
6.4.2	Mathematical Model and Ontology . . . . .	106
6.4.3	Implementing Proposed Mathematical model for Multimodal Integration in MagicHood . . . . .	109
6.5	Evaluating the Effectiveness of Proposed MagicHood Model: Empirical Findings and Discussions . . . . .	114
6.6	Conclusion and Limitations . . . . .	119
	<b>Part 5 - The Next Steps</b>	<b>121</b>
	<b>7 Conclusions and Future Directions</b>	<b>123</b>
	<b>8 Appendices</b>	<b>127</b>
	<b>Appendix A - GFRS inventory (VOLTAGE)</b>	<b>130</b>
	<b>Appendix B - Visual methods of communication (MagicHood)</b>	<b>131</b>
	<b>Appendix C - Survey forms for data collection</b>	<b>134</b>
	<b>Appendix D - Readability scores used</b>	<b>135</b>
	<b>Bibliography</b>	<b>135</b>
	<b>Publications of the Candidate</b>	<b>147</b>
	<b>Brief Biography of the Candidate</b>	<b>149</b>
	<b>Brief Biography of the Supervisor</b>	<b>151</b>
	<b>Brief Biography of the Co-Supervisor from BITS, Pilani</b>	<b>153</b>
	<b>Brief Biography of the Co-Supervisor at Infosys</b>	<b>155</b>



# 1 | Introduction

*“The art and science of asking questions is the source of all knowledge.”*

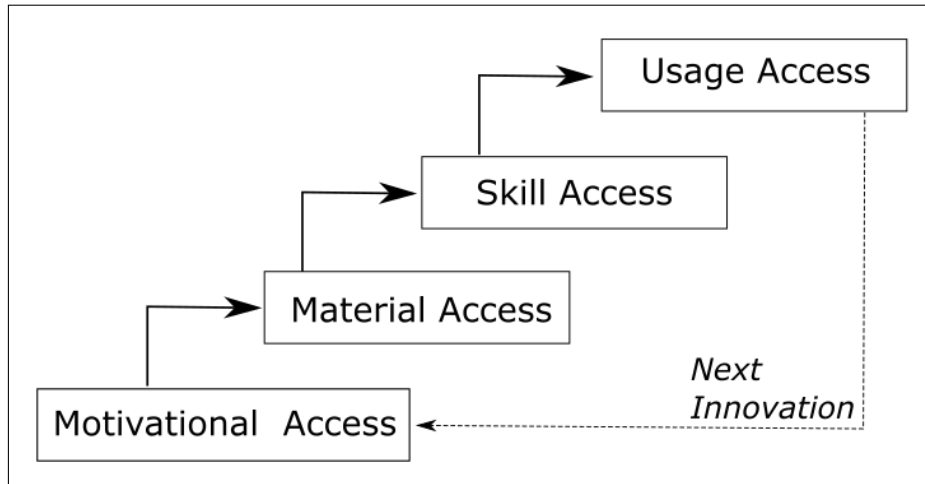
---

Thomas Berger

In the 21st century, the rapid advancement of technology has transformed the way we live, work, communicate and collaborate with each other. While the digital age has brought unprecedented opportunities, however in the developing nations it has also amplified the digital divide, especially for people with limited resources (reduced access to technology, low socioeconomic status along with diminished educational attainment). The issue of digital literacy becomes increasingly pertinent as technology permeates almost every aspect of human lives. This crucial imperative of digital literacy forms the basis of our research which we explore in a systematic manner and build AI-powered tools and techniques to facilitate digital adoption for people with limited resources.

While the investigations on the phenomenon of digital-divide shows that the classical social, psychological and cultural backgrounds would serve as the background, however language barriers, academic education and lack of digital infrastructure (both physical and usability) are equally important factors leading to the exclusion from the digital sphere. The objective of our research is to facilitate digital access applying the principles of Artificial Intelligence (AI), Machine Learning (ML), Natural Language Processing (NLP) along with exploring the state-of-the-art Generative AI (Gen-AI) methods and Large Language Models (LLM). We also overlay our AI-models with the principles from linguistic theories of semantic simplification like Natural Semantic Metalanguage (NSM) to further enrich our research.

Studies on the digital divide make observations pertinent to social, mental and technological causes and categorize the users issues into four levels of hierarchy [1]. As illustrated in Figure 1.1 the four possible challenges in technology adoption are illustrated below [2].



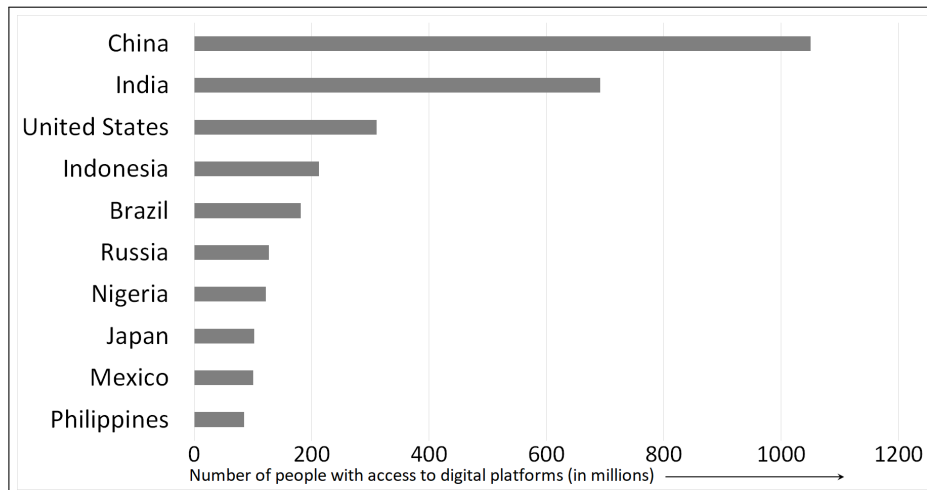
**Figure 1.1:** A cumulative and recursive model of successive kinds of access to digital technologies.

1. Lack of **Motivational-Access**, refers to the scenario where the end users do not have the wish to connect to digital world. This may also be referred to as '*want-nots*' and is generally studied in depth as part of humanities research.
2. Insufficient **Material-Access**, refers to the scarcity of physical access to personal computers/smartphones and the internet which boils down to demographic categories (income, education, location etc.) and can be referred to as '*have-nots*'.
3. Reduced **Skill-Access**, refers to the lack (or gaps) of skills to use digital devices. This primarily comprises of digital skills along with exposure to toolkits and navigating the internet. We can refer to this as '*un-skilled*'.
4. Limited **Usage-Access**, refers to the non-optimum use of digital ecosystem including limited use of internet for diverse applications and creative use. We can refer to this as '*in-active*'.

It is observed that the access types (motivational, material and skill) listed above as #1, #2 and #3 makes the technology accessible, it is the usage access (listed #4 above) that makes the technology engageable. Adoption (which will happen only when the technology is both accessible and engageable) of new technologies like digital communication shall eventually help in the improvement of conditions (economic, social and cultural) and enhanced quality of life.

The subjects of interest in our work comprises of under-represented, marginalised, academically impoverished and digitally challenged individuals. We architect socio-linguistic AI-models to make accessible and engageable digital technologies (addressing skill-access and usage-access) available for these users. We design and



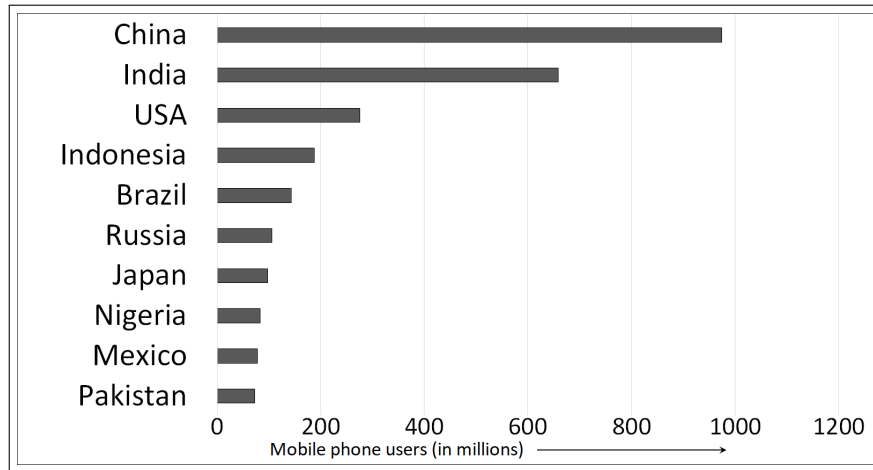


**Figure 1.2:** Adoption of digital platforms geographically

develop novel methods using the principles on AI, ML NLP, Gen-AI along with linguistic theories including Natural Semantic Metalanguage (NSM) and facilitate accessible and engageable digital information.

We contextualize our work with an Indian context (languages and demographics) to discover and define the scope of work, and design a variety of socio-linguistic AI-models to solve these problems in diverse circumstances. Keeping the hierarchical access model as the base, and assuming that motivational and material access is achieved, there are possible three categories of user types who needs assistance for digital enablement.

- **Skilled but inactive** users possess the necessary knowledge to operate digital devices, but they encounter difficulties due to low levels of engagement. These users require assistance on simplifying digital content to enhance accessibility and ease of use. The support needed for this group is minimal, as their main barrier is not skill, but re-engagement with digital platforms.
- **Partially skilled and in-active** users encounter engagement challenges due to limited familiarity with digital languages. Although proficient in their native language, their low motivation to engage with non-native digital content restricts their exploration of the digital ecosystem. Bringing these users to digital ecosystem needs digital on-boarding of low and ultra-low resource languages.
- **Un-skilled and in-active** face maximum hurdles in the use of digital platforms. These users use smartphones only for voice communication and they don't foresee any value to get skilled and explore more advanced use of digital technologies. These users face learnability issues and get quickly frustrated



**Figure 1.3:** Adoption of mobile phones geographically (in millions)

with adoption to new technologies. As a result, they require the most assistance in being on-boarded to digital platforms.

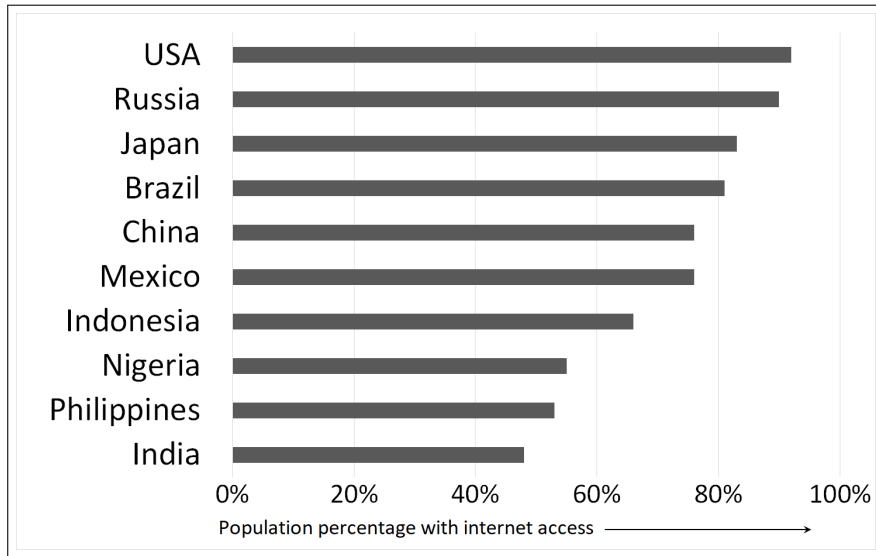
We explore each of these categories separately and address their needs to enable and successfully onboard them to the digital ecosystem leveraging the power of digital platforms.

## 1.1 Motivation

In order to substantiate the business case for this work along with the clearer understanding of the problem's scope and identify gaps we have leveraged data from multiple sources including "Our World in Data", "Statista", "World Bank", "Gapminder", "DataReportal" and "HYDE" [3, 4, 5, 6, 7, 8]. All these sources stand out as the ideal source for initial research and understanding the problem quantitatively, due to its unparalleled breadth and depth of data across demographics, its reliability sourced from reputable organizations, and its user-friendly interface, which enables swift access to comprehensive insights. The numbers, statistics, charts and other info-graphics in this chapter (Figure 1.2 - Figure 1.10) use these data catalogues as the sources for data.

With over two-thirds of the world's population already on-boarded on various digital platforms, the distribution of these users across geographies and languages is not consistent. We conduct Exploratory Data Analysis (EDA) as a crucial preliminary step to explore and understand the key features, trends and takeaways and further discover and define the initial hypotheses keeping the human aspect at the epicenter.

## The Business case



**Figure 1.4:** The reach of digital platforms across various countries.

With a global internet user base of 5.3 billion, amounting to 65.7 percent of the total world population there are still more than one third who have not signed on the web. A large percentage (61.4 % amounting to 4.95 billion) of the internet user base are only social media users. This clearly indicates that the use of digital platforms for societal good has a lot of scope for improvement and needs to be studied in detail.

### **The demographic divide**

As illustrated in Figure 1.2 and 1.3, the geographic digital spread and mobile phone reach (in absolute numbers) is more inclined towards Asian regions. It is not hard to join the dots that most of this is due to the population distribution and hence countries like India and China top the list. A close look at the penetration of internet (ratio of people with digital access/without digital access) in these countries, as illustrated in Figure 1.4 gives us an insight that the consumption (along with generation) of digital information will become more Asia centric in the near future. Within India, the digital penetration is less than half which means that as more and more people within India adopt digital methods of communication it can prove to be a great tool for operational efficiency, and improve the quality of life specially for people with reduced resources.

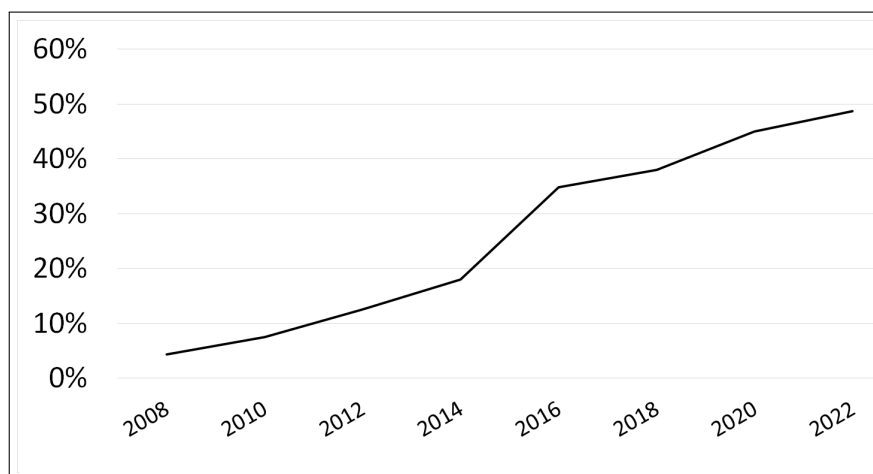
As Illustrated in Figure 1.5 we also study the growth of internet users over the years within India and find it to be steadily increasing. It has been observed that the initial growth primarily gets driven by motivational and material access [9]. In India, this has been possible by campaigns from the government to make people aware of the benefits from digital adoption along with providing financially viable options for mobile phones (devices). Though the campaigns help with the initial growth, however it has been observed that the adoption hits a plateau and achieving

full digital-literacy needs the careful study on the challenges faced by the users and build tools and techniques to address their challenges [9, 10, 2].

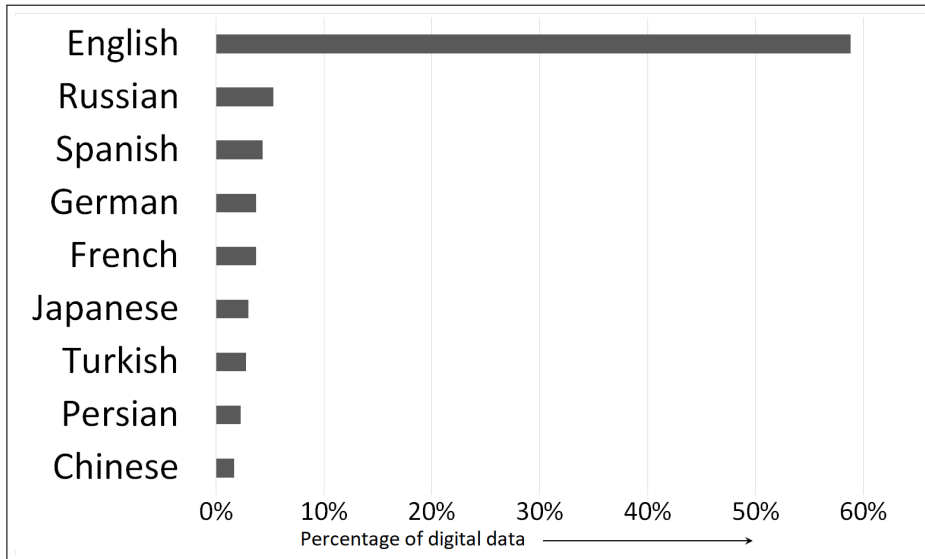
### The linguistic divide

Another perspective to analyse the digital-divide can be on the linguistic basis. The vast amount of information available on the internet in the form of web pages, social media posts, emails, online articles, blogs, forums, and other digital content primarily comprises of linguistic content (text, voice, videos etc.). As illustrated in Figure 1.6 and 1.7 there is a huge gap between the languages being spoken by people on the globe and the languages on the internet. English seems to be the most widely used and accepted digital language (digital lingua franca) for multiple disciplines, however people with lower levels of academic literacy (along with people who don't use English as their first language) face challenges to effectively consume the digital content in current form. It would be a good idea to diversify the digital content linguistically and enable digital on-boarding for more languages and scripts to facilitate the digital adoption for non-english speakers.

With over 7000 languages worldwide, and some 2000 within India itself the study on linguistic preferences across various states and cities in India can become a complex interpretation [12, 13]. There are studies on the preferred mother tongue of people within India which indicate that there are more than 60 languages with user base of more than 1 million. This is also illustrated in Figure 1.8 showing languages with more than 30 million user vase. It is evident that linguistic diversity in India is very broad and hence building language specific AI-models may not a scalable solution. It is important to conduct focused studies to address this challenge and design language agnostic AI-models facilitating digital on-boarding for marginalised languages.



**Figure 1.5:** Internet growth in India, in the last 15 years.



**Figure 1.6:** Linguistic distribution of web content.

### Accessibility barrier due to user interface design

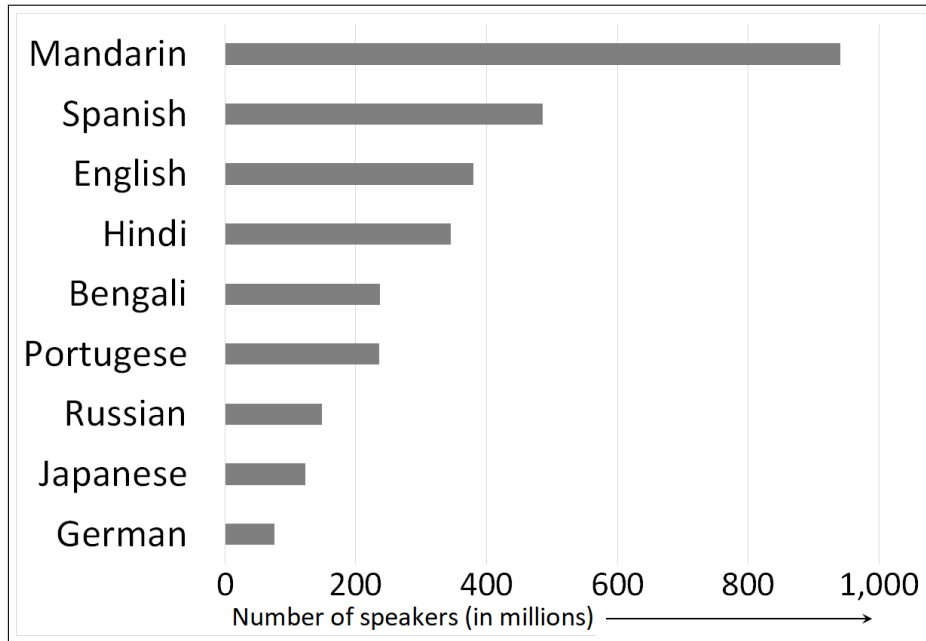
Effective UI design is crucial for enhancing digital literacy, especially for users with limited academic literacy and digital skills. There are studies conducted and observed that the traditional text-heavy interfaces can be challenging for low-literate users, emphasizing the need for simplified, intuitive designs. By incorporating visual aids along with easy-to-understand icons, digital technologies can become more accessible [14]. Having an intuitive user interface not only helps with improved user experience but also helps to improve the attention span of users.

Extensive field studies have been conducted and a set of design principles have been arrived including (a) Avoid text/ simplify text; (b) Increase photo-realism by incorporating abstracted graphics; (c) pay attention to psychological, cultural, or religious biases [15].

The arguments established herein clearly illustrates that there is a digital divide in the world (demographic and linguistic) and there is a clear business case to leverage technology and state-of-the-art AI-models to enable and facilitate digital-adoption for the users not yet on digital platforms.

## 1.2 Problem Discovery and Design Principles

It is very critical to deeply understand and frame the problem space before ideating the potential solutions. Understanding and empathising with our population of interest can provide insights on the needs and wants of end users and frame the problem space effectively. The proposed AI-models should not just meet users' needs but also deliver meaningful value to them.



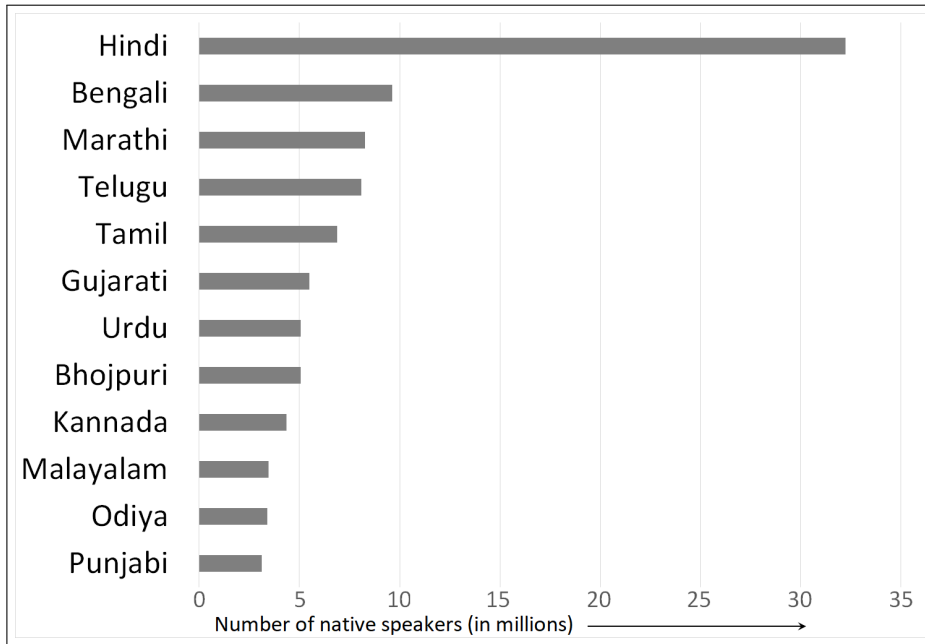
**Figure 1.7:** Most widely spoken languages worldwide [11].

## User Insights and Problem Discovery

In order to propose an effective, viable and desirable technology driven solutions, it is crucial to keep the perspective of end-users beyond our perception of their problem definition. It helps in understanding their definition of problems faced, circumstances and situation which helps in precise problem discovery and defining the problem space.

Center for the Advanced Study of India (CASI), conducted a survey to investigate the key aspects for lower digital literacy in India. Their hypothesis was to conduct a survey where they probed the respondents (randomly selected across urban and rural demographics) by asking a very fundamental question *In the developing world, where digital adoption is a challenge, what is your comfort in comprehending the language of the internet.* As illustrated in Figure 1.9, the findings from this survey are categorised on 4 parameters (demographics, economics, education and social status) and it is noteworthy that all of these four parameters play an important role in the adoption of the digital content (since digital content is mostly in English language). Some of these parameters in itself would have a strong correlation, for example people with more financial prosperity will have better access to schools and universities and hence would be more educated.

Ministry of Women and Child Development, Government of India along with University of Delhi conducted a similar study on the challenges with the digital literacy in India, more specifically for urban poor women. The findings from their research also indicates that tertiary education is one of the factors for being digitally

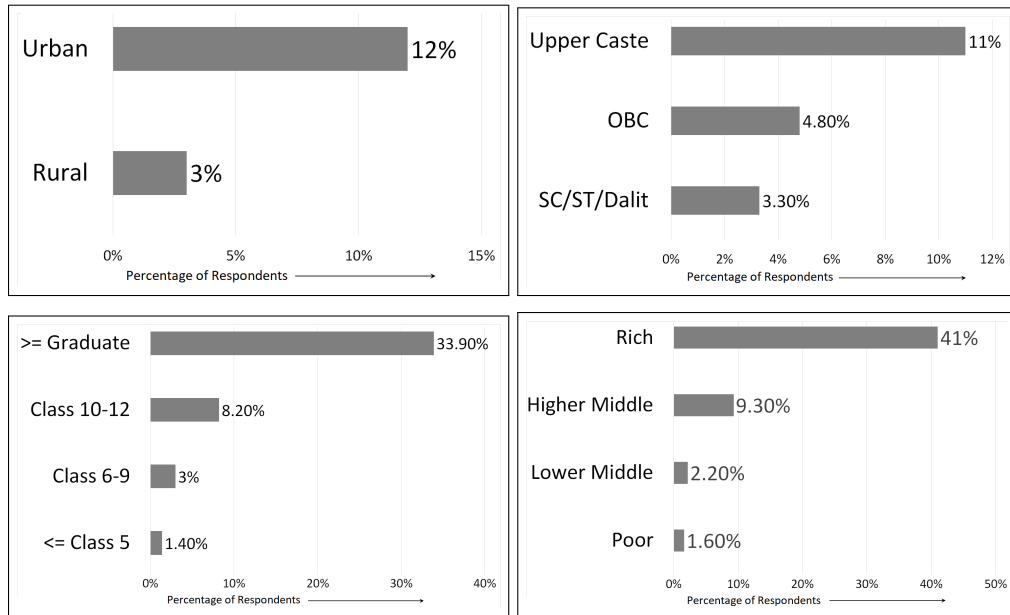


**Figure 1.8:** Native languages in India with more than 30 million users.

illiterate [16].

Further more deeper studies conducted by India Human Development Survey (IHDS) and Annual Status of Education Report (ASER) it is evident that (a) Urban areas generally have better educational infrastructure, more qualified teachers, and better access to supplementary educational resources compared to rural areas; (b) Economically prosperous families are more likely to afford private schooling and higher education, leading to better educational outcomes; (c) Historical and social factors have resulted in higher castes having better access to education. Affirmative action policies like reservations have improved access for lower castes, but disparities still exist. Concluding to these observations and findings, we consider academic education as the key qualifier (which is anyway correlated to other factors) responsible for digital-illiteracy and define our population of interest on education [17, 18].

We qualify people with academic literacy less than 12 years of formal schooling, due to restricted schooling facing challenges with the use of digital technologies, as our targeted-users. We refer these people as **semi-literates** in our work, and observe that while majority of these people would have surpassed motivational and material access, however they still face challenges with skill and user access. In order to gain ground-zero insights, we conduct an in-depth survey as a part of our work reaching out to these users and understand their problems using data driven findings. Following the findings from the survey, we categorise this groups of people into smaller subgroups for focused attention. We illustrate the details from our survey in detail as part of Chapter 2.



**Figure 1.9:** Familiarity with English across (a) Demographics (b) Social discrimination (c) Education and (d) Economic prosperity

### Challenges faced by semi-literates

In summary, the challenges and problems faced by digitally-illiterate communities as concluded with existing studies and our survey observations can be broadly classified into two types.

- **Language (on the digital platforms) Exposure:** Most semi-literate users struggle with the understanding of complex words on the digital platforms leading to confusion and misinterpretation.
- **User Experience and Limited Attention Span:** Lack of intuitive navigation and clear instructions which can help these users to access the digital platforms. The cognitive load associated with comprehension renders most semi-literate individuals unable to effectively utilize digital technology.

### Research Maxims

After a careful analysis of the observations, challenges and problems faced by our subjects of interest, the next step is to clearly define the purpose of our study which also outlines the scope of our work. This helps to establish a concise idea of what exactly we are trying to solve and how relevant it shall be in solving the problems faced by end users, the broader articulation of the objectives. These goals acts as a north star for the rest of the design process. We outline three key objectives/design-principles for our work.



- **Simplicity:** Proposed AI-models should be easy to use, have a reduced learning curve, facilitate intuitiveness and reduce ambiguity.
- **Extensibility:** Proposed solutions should be provisioned for scalability, language agnostic adoption and able to extended to multiple domains.
- **Digital-First:** Proposed AI-models should embrace digitally native and optimised for digital ecosystem.

## 1.3 A Comprehensive Literature Review for Enhancing Accessibility and Engagement in Lower Education Cohorts

Making digital enablement accessible to everyone involves addressing various factors to ensure equitable access and usability. For the context as outlined in our problem statement and subjects of interest, it has been observed that the earlier studies have made attempts to build accessible technologies thereby increasing digital-literacy primarily using three approaches, (a) Semantic simplification of the digital content; (b) Facilitating digital communication in native languages (along with localised scripts) addressing language barriers; and (c) Visual communication methods for users with bare minimum academic education, minimising the use of orthographic text. We discuss each of these briefly here and elaborate further in individual chapters where we apply some of these design principles as part of our work.

### 1.3.1 Text Simplification: Enhancing Comprehensibility along with preserving Semantic Integrity

Text simplification aims to make written forms of the languages more accessible and understandable, particularly for individuals with lower academic education, language learners (who use that language as a second language), or those who may struggle with complex vocabulary. Text simplification is a mix of theoretical linguistic approaches along with computational methods by the use of deep learning frameworks to achieve reduction in complexity and easy comprehensibility. In our work, we achieve lexical, semantic and morphological simplification applying a variety of linguistic and computational approaches.

### Natural Semantic Metalanguage (NSM)

The Theory of Natural Semantic Metalanguage (NSM) is a linguistic theory developed by Anna Wierzbicka and her colleagues. It proposes that there is a set of universal semantic primitives, or "semantic universals" which are the building blocks of all human languages. These primes are simple, irreducible meanings that are common to all languages and cultures [19, 20].

NSM aims to simplify language by breaking down complex concepts into their simplest components using semantic universals. NSM seeks to make language more transparent, understandable, and cross-culturally applicable. NSM can enhance communication, improve comprehension, and promote inclusivity in digital content primarily via reduction of complex vocabulary and cross-cultural communication.

In our work, we have applied NSM at multiple junctures. In the process of text simplifications NSM helps with simplification of structure and formation of words. We also leverage NSM for pictographic communication as an alternative form of communication.

## Paraphrasing

Paraphrasing consists of different linguistic forms (lexical, semantic, syntactic, morphological, discourse) expressing the same meaning. While the definition of paraphrasing requires strict semantic equivalence, linguistics accepts a broader, approximate, equivalence also termed as "quasi-paraphrase". Paraphrasing originated as a linguistic phenomena, and later became a point of interest by computational researchers. Table 1.1 illustrates high level paraphrase typology along with various types and examples [21].

Paraphrasing plays a crucial role in semantic simplification, which involves conveying complex information in a clearer and more accessible manner. In this digital world, paraphrasing can be used to simplify technical language, complex concepts, and jargons to enhance readability and understanding for diverse audiences. By rephrasing information in simpler terms while retaining the original meaning, paraphrasing helps to make the content more accessible to readers who may not be familiar with specialized terminology or enhanced vocabulary.

Computationally the task of paraphrasing can be divided into three separate categories namely, (i) Paraphrase Recognition (PR), (ii) Paraphrase Generation (PG) and (iii) Paraphrase Extraction (PE). Based on the implementation approaches for these sub tasks, sometimes these may overlap and sometimes they are independent of each other. In order to achieve semantic simplification, PE is generally used to create a parallel-corpus of complex and simple phrases which can be used to train a deep learning PG models. PE is used to empirically validate the quality of the models architected.

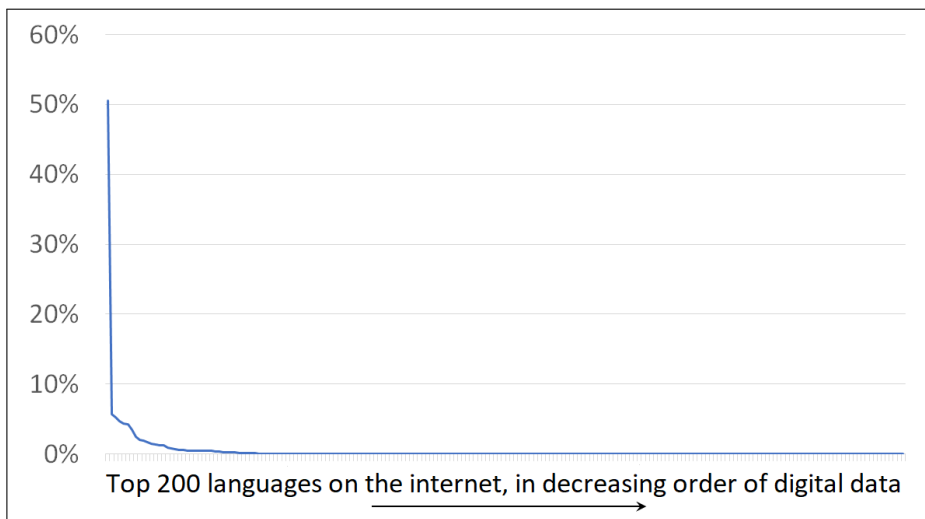
From the functional perspective, the phenomena can be categorised into three categories, (i) Paraphrase Entailment, where for the given phrases  $p_1$  and  $p_2$ , if the phrase  $p_1$  is to be trusted then it would infer that  $p_2$  is most likely also true. Entailment may not be always bidirectional in nature [22]; (ii) Conditional Paraphrase, where additional constraints (morphology based, syntax-based, readability-based etc.) are applied while paraphrasing is conducted [23]; (iii) Reductive Paraphrase where complex language concept is paraphrased into a set of lexicons which are easy to interpret. This category of reductive paraphrasing is most relevant for our work.

**Table 1.1: Paraphrase typology for various types**

Paraphrasing Typology	Examples
<i>Lexicon Based changes for paraphrasing</i>	
Same polarity substitution	(1) a. Google bought YouTube b. Google acquired YouTube
	(2) a. They were 9 b. They were around 10
	(3) a. The pilot took off despite the stormy weather b. The plane took off despite the stormy weather
Opposite polarity substitution	(4) a. I am leaving b. I am not staying
	(5) a. Only 20% of the participants arrived on time b. Most of the participants arrived late
	(6) a. Google bought YouTube b. YouTube was sold to Google
Deletion	(7) a. I like eating chocolate b. I like chocolate
	(8) a. Actually, you shouldn't be here b. You shouldn't be here
Synthetic/analytic substitution	(9) a. Steven attempted to stop playing Hearts b. Steven made an attempt to stop playing Hearts
	(10) a. Ideas is all I need to write an article b. A sequence of ideas is all I need to write an article
	(11) a. I prefer wildlife television documentaries b. I prefer television documentaries about wildlife
<i>Morphology Based changes for paraphrasing</i>	

Inflectional change	(12) a. In 1492 Columbus reached America b. In 1492 Columbus reaches America
Derivational change	(13) a. I know that Olds founded GM b. I know about the foundation of GM by Olds
<i>Syntax Based changes for paraphrasing</i>	
Diathesis alternation	(14) a. John loves Mary b. Mary is loved by John
	(15) a. The laundry sways in the breeze b. The breeze makes the laundry sways
	(16) a. The section chief filled Japanese sake into the cup b. The section chief filled the cup with Japanese sake
<i>Semantics Based changes for paraphrasing</i>	
Lexicalization pattern change	(14) a. Bill flew across the ocean b. Bill crossed the ocean by plane
	(15) a. The increase of prices accompanies the crise b. The prices increasent with the crise
	(16) a. Barbara excels at teaching b. Barbara teaches well
<i>Discourse Based changes for paraphrasing</i>	
Discourse structure change	(14) a. He wanted to eat nothing but apples b. All he wanted to eat were apples
	(15) a. Joe wants the blazer which was designed by BMW b. Joe wants the blazer designed by BMW
	(16) a. He is willing to leave. This made Gillian upset b. His willingness to leave made Gillian upset

A variety of computational methods can be leveraged for paraphrase generation including, Supervised Monolingual Machine Translation (SMMT) which includes Statistical machine translation (SMT), Neural machine translation (NMT), Pivot based Machine Translation, Unsupervised Machine Translation etc [24, 25, 26, 27]. The use of Re-enforcement Learning is also getting popular having a generator (for paraphrase generation) and evaluator (for paraphrase evaluation) working



**Figure 1.10:** The long-tail of linguistic distribution on the internet.

together as a re-enforced learning system similar to the configuration of Generative Adversarial Networks (GAN) [28, 29].

In our work we leverage transformer based neural machine translation thereby achieving reductive paraphrasing for text simplification and making it relevant for semi-literate populations.

### 1.3.2 State-of-the-Art Techniques for Low Resource Languages

Modern NLP research focuses primarily on the languages on the internet, which consists of only 20-30 of the 7000 languages of the world [12]. This leaves the majority of languages understudied, which are also referred to as low-resource languages (LRLs), and are widely spoken by a large section of world population. While we have illustrated the linguistic distribution of the internet as part of Section 1.1, however this is further illustrated in Figure 1.10 where top 200 languages on the internet have been plotted with the volume of digital content for them. This illustration is generally referred to as the long-tail problem on linguistic diversity of the internet and needs to be addressed for an inclusive internet.

LRLs can be described as *resource scarce, under studied, less digitized, under privileged or less commonly taught*, among other denominations [30, 31]. For most LRL there is not enough digital data available and hence NLP methods cannot be directly applied, since most new methods of computational linguistics are data hungry. There are more than 2.5 billion inhabitants using 2000 LRLs, within India and Africa itself and any progress for these languages shall help in digital enablement of these

**Table 1.2:** Language Resource Distribution with number of languages and speakers. Category 0,1, 2 and 3 are referred to as Low Resource Languages (LRL)

Cat.	# of Languages	Speakers	Labelled Data	Unlabelled Data
0	2191	1.2B	Very Low	Very Low
1	222	30M	Very Low	Low
2	19	5.7M	Low	Medium
3	28	1.8B	Medium	Medium
4	18	2.2B	High	High
5	7	2.5B	Very High	Very High

populations. Table 1.2 illustrates the classification of various languages analysing the digital status and ‘richness’ in the context of useful linguistic resources (labelled and unlabelled) and number of speakers [32].

The most fundamental and critical NLP task to onboard the low and ultra-low resource languages comprises of (i) Design of Optical Character Recognition (OCR) methodology for the underlying script(s) (may also involve creation of digital fonts and styles) for low-to-no-data scenarios, and (ii) Design of AI-models to create/ augment bilingual data corpus to be used for state-of-the-art deep learning methods. Only once these tasks are accomplished, most advanced NLP tasks including translation/ transliteration, sentiment/emotion analysis, named entity recognition, summarizing etc. can be designed and engineered.

## Optical Character Recognition (OCR)

Though the early ideas of OCR dates back to 1870’s, however the initial studies were focused more on specific languages and special hardware and limited to high end labs only [33]. Even in the current times, the scope of OCR research is restricted to high resource languages or those languages where enough digital content is available. Within India, the major work in Indic OCR is limited to the ten scripts namely, Bangla, Devanagari, Gurmukhi, Gujarati, Kannada, Malayalam, Oriya, Tamil, Telugu, and Urdu [34]. Most OCR methods are architected using deep neural networks which tends to be hungry on data and computational power. There are a few off-the-shelf OCR systems available today (both open source and licensed) including Tesseract, OCRopus, Kraken and Calamari [35, 36, 37, 38] etc.

OCR pipeline generally goes through multiple individual tasks including (a) Image acquisition (extracting images containing text from multiple sources for offline images, and capturing live images for online extraction) (b) Pre-processing (application of image processing techniques, to increase raw image quality) (b) Binarization

(for scenarios where text and images/videos are mixed, we need to isolate text images from background) (c) Layout Analysis (dividing the images into regions) (d) Segmentation (segmentation of image into pages, lines, words, characters and symbols) (e) Feature Analysis (identification and extraction of key features) (f) Classification (Recognition of symbol with scrip character-set) (g) Post processing (use of pre-compiled vocabulary and language rules to auto correct the unrecognized words) [39]. To the best of our knowledge there is no general purpose OCR method for ultra low resource scripts where there is little to none digital data available.

We have worked on a variety of ultra low resource scripts (Takri, Modi, Ol Chiki, Wancho) in our work and designed a novel OCR method which can work on low-to-no-data scripts.

## Data Augmentation (DA)

Data Augmentation (DA) refers to the methods that aim to expand the diversity and volume of data. While DA has become a standard technique to enrich data for deep networks in NLP tasks, it is still not a common practice in Low Resource Languages (LRL) and scripts, where there is dearth of digital data.

DA schemes are classified into two parts, extractive and generative. While *Extractive approach* for DA uses unsupervised data from multiple sources, applying web crawling, voice transcriptions, document scanning etc., *Generative approach* uses artificially synthesised data using various kinds of text synthesis approaches [40]. Extractive DA can be achieved via (a) Rule-based technique (b) Interpolation technique and (c) Model-based techniques [41]. Generative DA on the other hand is achieved by (a) Paraphrasing (b) Noising and (c) Sampling based methods [42, 25].

DA to accomplish the specific task of bilingual data corpus is of paramount importance for low-high resource language combination. Such corpora can serve as invaluable resource for developing and training AI-models, for tasks such as machine translation, cross-lingual information retrieval, and sentiment analysis across different LRLs. This can also enable to effectively process and understand various linguistic nuances and cultural contexts. Bilingual corpora plays a crucial role in bridging language barriers, fostering cross-cultural understanding and promoting linguistic diversity and inclusivity in the digital age.

We have leveraged all these DA techniques to create an algorithmic bilingual data corpus generation method for low resource languages, which uses newspaper articles as raw data.

### 1.3.3 Innovations in Visual Communication Methods for Enhancing Accessibility in Low-Education Populations

Visual methods of communication can greatly help as communication methods for people with lower literacy, cross cultural communication or people with learning disabilities. Visual methods of communication has been a topic of interest both for linguists and computer scientists and the approaches for both have been different. In the recent times there are also studies with inter-disciplinary that integrates concepts, methods, and perspectives from both linguistic and computational perspectives.

#### Linguistic Methods

The use of pictographs has been in use for written communication from very ancient times and can be seen inscribed in various old monuments. However, most of the scripts lost its popularity with the advent of orthographic and phonetic scripts which required training and education to interpret and comprehend [43]. There have also been some attempts on building large pictographic/ideographic dictionaries using volunteer contributions to create large inventories of pictographic lexicons like PictNet and Noun-Project<sup>1</sup> [44].

One of the early pictographic scripts which is still in use and being researched by Chinese and Tibetan researches is Naxi. This script belongs to the Yi language branch of Tibetan-Burmese languages and used in the Yunnan province of southwestern China [45].

International System of Typographic Picture Education (ISOTYPE) was developed by Austrian philosopher and social scientist Otto Neurath and his wife Marie Neurath in the 1920s and 1930s as a method of visual communication. It uses a pictorial form within a two-dimensional syntax to show social, technological, biological and historical connections [46].

Blissymbolics is a symbolic language system designed to facilitate communication for individuals with severe communication impairments, such as those with developmental disabilities or conditions like cerebral palsy. The vocabulary has 100 basic symbols, and has some combinational technique to create more words [47].

Nobel Language named after the chemist Alfred Nobel, is a pictographic language based on 120 basic signs and many arrows of different shapes to facilitate communication between speakers who do not share a common language. This was

---

<sup>1</sup><https://thenounproject.com>



developed at the department of Mathematics and Computer Science at Drake University [48].

Pictographic methods as illustrated are a few pioneering studies done by linguists and social scientists in the field of visual communication. Nonetheless, the concept of using pictographic languages as a universal means of communication has been explored by various individuals and organizations throughout history.

## Computational Methods

Iconji is a pictographic language or communication system developed by Omnipotent Media, designed to serve as a universal language for digital platforms. It encompasses an open, visual vocabulary of characters with built-in translations to twelve languages. Unfortunately, the project was put on hold due to lower user engagement [49].

"Able to Include" is a specialised project focused to improve the living conditions of people with intellectual or developmental disabilities (IDD) using Augmentative and Alternative Communication (AAC). It uses Scalera and Beta as pictographic scripts which are enabled for digital platforms and communication [50, 51].

Text to Picture (TTP) synthesis is a more generic text to picture synthesis system and comprises of two fundamental steps namely (i) Identification of key elements which can be converted to pictures and (ii) Selecting images to convert those key elements to text. In the recent times, TTP synthesis is continues to evolve with advancements in Generative-AI (Gen-AI) and Machine Learning (ML) techniques. As the technology improves, it holds the potential to enhance various aspects of visual communication and content creation. We mention the use of Gen-AI separately in next section.

## Generative AI

In the recent years, the rapid pace on the development of Large Language Models (LLM) and public release of tools such as ChatGPT/ LLAMA has attracted wide attention, optimism and concerns [52, 53, 54]. These models are capable of generating new content, in forms of text, audio, visuals etc. understanding the patterns in existing data, and hence is referred to as "Generative-AI".

LLMs have become the backbone for most NLP tasks primarily "text generation" however can also be leverages for a variety of NLP tasks including text understanding/inferencing and classification tasks as well [22]. LLMs encapsulate world's knowledge in form of parametric memory and can be very instrumental in referencing new concepts not seen in model training. In order to maximise the effectiveness of LLMs, a carefully crafted textual phrase, also referred to as "prompt" is provided

to generate appropriate output. The prompt can contain, instructions, context and any other constraints to fine tune and get appropriate results.

We leverage all of these techniques (linguistic, computational and GenAI) and design a novel multimodal communication metalanguage for semi-literates which delivers information to the end user with reasonable accuracy.

## 1.4 Research Objectives, Scope, and Boundaries

### Gaps and Objectives

Though there has been significant work done on digital enablement for a variety of users (facing challenges with digital technologies), using a variety of computational methods, however end-to-end AI enabled models and design of pipelines for semi-literates is still far fetched and requires focused work. There is still a large gap to innovate and design accessible and engageable AI-models which can bridge the digital divide and build an inclusive digital world.

It is also noteworthy that the asks and challenges of every semi-literate individual facing difficulty to get on-boarded to digital platforms, may not all the same for everyone. It therefore becomes pivotal to partition the broader group of semi-literates into specific categories based on the challenges and expectation of each sub-group and outline the broad objectives as under.

#### 1. **Semi-Literate Category:** Digitally-Novice

##### **Gaps Identified:**

- Language Exposure: People from this group have the basic know-how to access and engage with digital platforms, and comprehend simple digital language constructs, however their exposure to complex vocabulary is a challenge.

##### **Research Objective:**

- Text simplification: Tools and techniques for this group would require innovative ways of building simplified digital ontology and the use of the same to design reductive paraphrasing systems in order to achieve reduced complexity of digital content.

#### 2. **Semi-Literate Category:** Digitally-Niche

##### **Gaps Identified:**

- Exposure to Digital Languages: People from this group are relatively well read however their choice of languages (and probably scripts as well) have no (or very minimal) digital footprint.

**Research Objective:**

- Onboard remote scripts to digital ecosystem: Propose general purpose OCR for marginalised scripts with little-to-nil digital-data.
- Curate datasets for low resource languages: Propose general purpose algorithmic methods to build good quality, large volume bilingual corpora.

**3. Semi-Literate Category: Digitally-Neglected****Gaps Identified:**

- Academic literacy and exposure to orthographic scripts: People from this category exhibit reluctance and disinterest towards digital adoption due to very low academic literacy, and their exposure to orthographic scripts is bare-minimum.

**Research Objective:**

- Design alternative methods of communication: Propose AI-models for this audience with the deeper understanding of their characteristics and preferences, and design methods of digital communication which minimizes the use of orthographic scripts.

## Scope

The scope for proposed AI-models is to design theoretical frameworks and methodologies, agnostic of languages, social and cultural behavior, and choice of hardware. The proposed AI-models needs to be validated on a variety of languages and scenarios along with comparisons with the state-of-the-art existing methods, to establish the merit of usage. While the underlying data to test/train AI-models mostly will come from Indian context (languages, user behavior, cultural preferences) and pre-trained models will be useful for Indian use, the proposed methods should be universal and easily configurable.

## Limitations

While the proposed AI-models shall facilitate on-boarding and engagement of our subjects of interest, the technologies should complement, rather than replace the human support and intervention in digital literacy initiatives. Additionally the data

used for training AI-models, should protect users' personal information (as per the regulatory needs) even if it leads to drop in model accuracy. Last but not the least, the training data in our work predominantly consists of samples from Indian demographics, hence the models in their current forms may not generalize well to universal populations.

## 1.5 Major Contributions

We propose AI-models for accessible and engageable digital communication for each category of semi-literates, thoroughly validated and empirically tested to establish our claims. We have also published our findings and observations for each proposed model in reputed conferences and journals.

- For *Digitally Novice Audience* we propose a novel zero-shot learning approach to achieve reductive paraphrasing for semi-literates achieving an improvement of an average of 25% in ease of comprehension.
- For *Digitally Niche Audience* we design two independent models, (i) a comprehensive unsupervised OCR methodology, VOLTAGE, for digital on-boarding of ultra low resource scripts, and (b) a novel methodology using image and text analytics to build a completely automated, scalable and language agnostic bilingual parallel dataset for low source scripts.
- For *Digitally Neglected Audience* we design a Generative AI powered multi-modal ideographic metalanguage as an innovative and alternative communication method that transcends the barriers of linguistic and cultural backgrounds achieving an accuracy of more than 80% on semantic comprehension.

## 1.6 Thesis Organization

We divide the rest of the thesis into 4 parts:

- Part 1 of our thesis which comprises of Chapter 2, establishes the groundwork for our analysis which forms the backbone for the proposed AI-models in later chapters. We conduct extensive fieldwork (collect text-messages and conduct interviews) and analyse the same on socio-linguistic patterns. We define the personas of semi-literates, for each category and articulate their challenges and asks in detail.

- 
- Part 2 of our thesis, focusing on the enablement of "digitally-novice" semi-literates comprising of Chapter 3 illustrates the proposed AI-models to achieve semantic simplification of complex English constructs using the late fusion transformer architecture, along with empirical validation of our model on qualitative and quantitative parameters.
  - Part 3 of our thesis, comprising of Chapter 4 and 5 explores multiple methods for "digitally-niche" semi-literate communities. While chapter 4 illustrates the design of OCR methodology for scripts with little-to-no-data scenarios, the chapter 5 illustrates a general purpose bilingual-dataset generation framework for low resource languages.
  - Part 4 of our thesis proposes AI-models for "digitally-neglected" groups. Chapter 6 illustrates a novel multimodal ideographic method of communication using ideographs powered by Generative AI, to enable digital communication for people with very low levels of academic literacy.

As a by-product, we also compile large datasets for public use as a part of our thesis. All proposed AI-models are empirically tested and published in peer-reviewed conferences/journals.



# **Part 1**

## **The Groundwork**





## 2 | A Versatile Dataset for Socio Linguistic Assessment of Semi Literate Urban and Rural Populations in India

*“The formulation of the problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill.”*

---

Albert Einstein

This chapter presents a comprehensive dataset that we have meticulously developed through our dedicated fieldwork efforts. The dataset is created by collecting information via interviews and text message data from the end users. The objective of the survey is to gain ground-zero insights into the actual problems and challenges of semi-literate individuals while using digital devices. We also leverage this data for validating our proposed AI models, in later Chapters.<sup>1</sup>

### 2.1 Semi-literate Texting (SLT): A dataset for the semi-literates, by the semi-literates

In light of the persistent digital divide and the vital play of technology within the modern society, it is imperative to undertake a comprehensive perspective of our subjects of interest, and use that data to derive insights.

Access to digital devices and the internet is often taken for granted in today’s interconnected world. However, a significant portion of the population, particularly

---

<sup>1</sup>Parts of this chapter have been published in Data in Brief, ELSEVIER, Volume 38, October 2021 [55]

those with limited academic education along with financial and social disadvantage, encounter challenges when attempting to utilize these technologies effectively. Therefore, it becomes crucial, to conduct an exhaustive fieldwork and conduct a survey to identify and understand the specific challenges faced by semi-literates, by the careful examination of text messages shared by these individuals and conducting interviews.

## **Defining semi-literates**

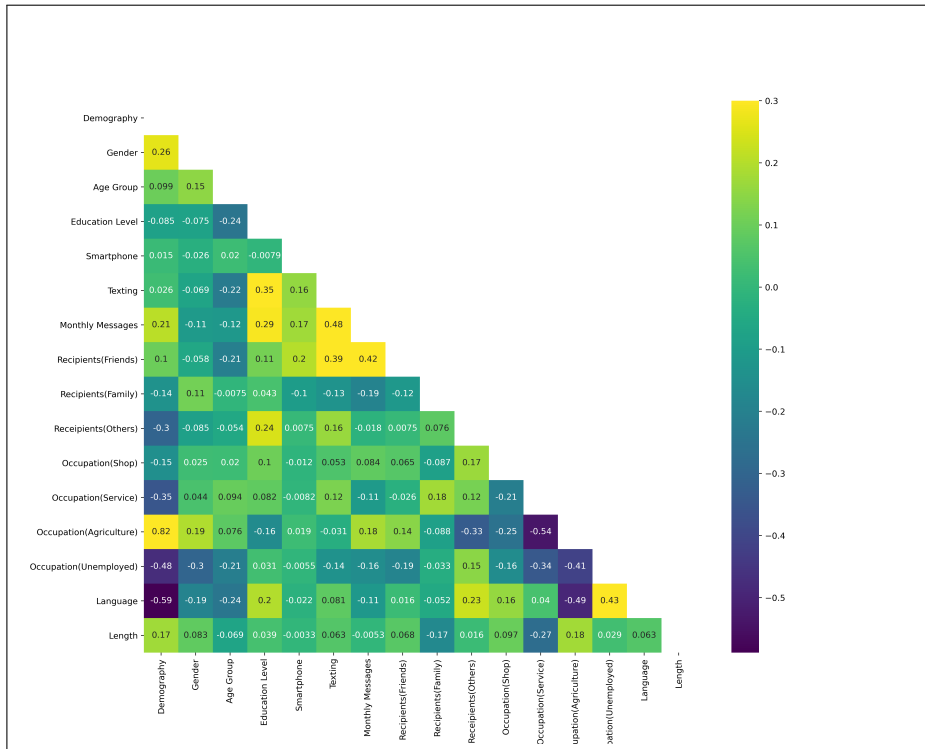
There is no clear definition of digital literacy in the research world. In late 1900s Gilster was the first one to coin the term ‘digital literacy’. He measured this on parameters of education and information skills [56]. Over a period, the definition of ‘digital literacy’ has modified. Chase and Laufenberg have referred this to be ‘inherently squishy’ [57]. The current definition ranges from being technology fluent to being able to use information on digital platforms without assistance.

In a recent study conducted by ‘Ministry of Women and Child Development, Government of India’ along with ‘University of Delhi’ it was observed that one of the key factors for being digitally illiterate is lack of tertiary education [58]. In another study a similar observation was made, that the poor literacy rate in India, particularly women and rural population is major impediment to the growth of digital literacy. “It is hard to think of universal digital literacy without universal academic literacy” [59]. Therefore it is evident from related literature that academic literacy is strongly co-related with digital literacy.

While most contemporary work classifies people as digitally literate or illiterate. We believe that there is another classification of users, who are digitally semi-literate. These users have access to mobile phones and internet, however face issues with the use of digital technologies due to lack of elementary education. We define semi-literates as those individuals who have spend less than twelve years in academic formal education and face challenges with digital devices.

## **Utilizing Text Message Data for Socio-Linguistic Analysis**

It is evident that the easy availability along with reasonably priced mobile phones and internet access in India has helped in penetration of this technology at grass-root level (motivational and material access, taken care of) [60, 18, 17]. However, most of the information on digital platforms which can possibly help emergent mobile phone users, is not suitable for their consumption due inability to navigate digital platforms and resources effectively [14, 15]. It is observed that the most practical application of digital technology for these users is the exchange of text messages



**Figure 2.1:** Heatmap for various parameters.

across their community, in order to communicate and participate in social, economic and political issues that could have an impact on them.

Most mobile phones facilitate textual communication which is asynchronous, can be sent and received quickly and more discreet than phone calls, allowing individuals to communicate privately in public settings or situations where speaking aloud may be disruptive or impractical. Semi-literates are conversant with local language and use the same for text communication. In most scenarios they use roman (English) script for text messages in native language, because of the ease of typing on QWERTY keyboards. It is also observed during our survey, that the use of non-native script creates challenges with comprehending information due to ambiguity and mismatch of phonemes.

## 2.2 SLT Dataset: Experimental Design, Materials and Methods

We conduct a detailed in-person survey, to study the text messages and associated metadata from digitally semi-literate mobile phone users in India <sup>2</sup>. A survey among urban and rural communities conducted between July 2020 and November 2020 is

<sup>2</sup><https://data.mendeley.com/datasets/4b53nj78tv/8>

the source for this dataset. The data has been collected through face to face interviews across urban and rural geographies in India, largely from western region of Maharashtra. A total of 382 respondents, accumulating 3368 messages has been composed (approximately 90% interviews are conducted face to face and the remaining 10% are conducted in online mode). Please refer to Appendix A, which illustrates the forms and the information collected as hard copies. The forms were manually collected and hand typed to create a digital repository and published [55].

We collect the actual text messages (after careful consideration of personal data) exchanged on digital platforms, by digitally semi-literate users and use the same to model semi-literate personas. This helps us to understand and decipher the sociolinguistic challenges of these users. Table 2.1 describes the specifications of the variables used during the data collection process. We observe that more than 70% of text messages exchanged, use regional languages (sometimes using roman scripts, sometimes native scripts). All the text messages have been verified for Personally Identifiable Information (PII) information and any PII content inside the messages has been anonymized.

We conduct a high-level statistical analysis on the survey data and present the summary statistics (impact of various parameters Length of text message) in Table 2.2. We also conduct correlation analysis across multiple parameters and represent the same as a heatmap illustrated in Figure 2.1. The purpose of heatmap is to identify patterns and correlations across multiple features collected which can help us with feature engineering later. As illustrated in Figure 2.2, we also conduct a high level sentiment analysis on this data along with generating the word cloud (across urban and rural participants separately) to understand the focus and sentiment of conversations [61].

The purpose of this exercise is to conduct preliminary exercise to understand if some parameters need some special considerations, like should we consider urban and rural users separately or not. And the results from this exercise indicated that barring only a few correlations (like people in rural areas use native languages more than urban participants) most behavioral statistics are similar and there is no need to split the dataset into partitions only on metadata. A more deeper analysis on textual messages (semantic, lexical, syntactic) is needed to understand and categorize the the users.

**Table 2.1:** Survey data details - variables and their description

<b>Variable</b>	<b>Type</b>	<b>Description</b>
Mode	Categorical	Face to face Online
Demography	Categorical	Rural Urban
Agency	Categorical	Agency Code
Gender	Categorical	Male, Female, Others
Age	Numeric	Age of the respondent
Town/Village, City, State	Numeric	Town/Village, City and State of residency
Education	Numeric	0 –No formal education 5, 8, 10, 12 –Highest level of education
Do you use Smartphone	Categorical	Yes, No
Do you send text message	Categorical	Yes, No When selected ‘No’ these recipients only receive and read messages
Frequency	Categorical	Daily, Weekly, Monthly, Never
Recipients	Categorical	Family, Friends, Employer, Others
Outstation communication	Categorical	Yes, No
Profession	Categorical	Categorical profession of interviewee
Language	Categorical	English, Hindi, Marathi, Others
Message (1-10)	Text	Translated English language text messages free from Personally identifiable information (PII) data
Length	Numeric	Average length of the translated English message



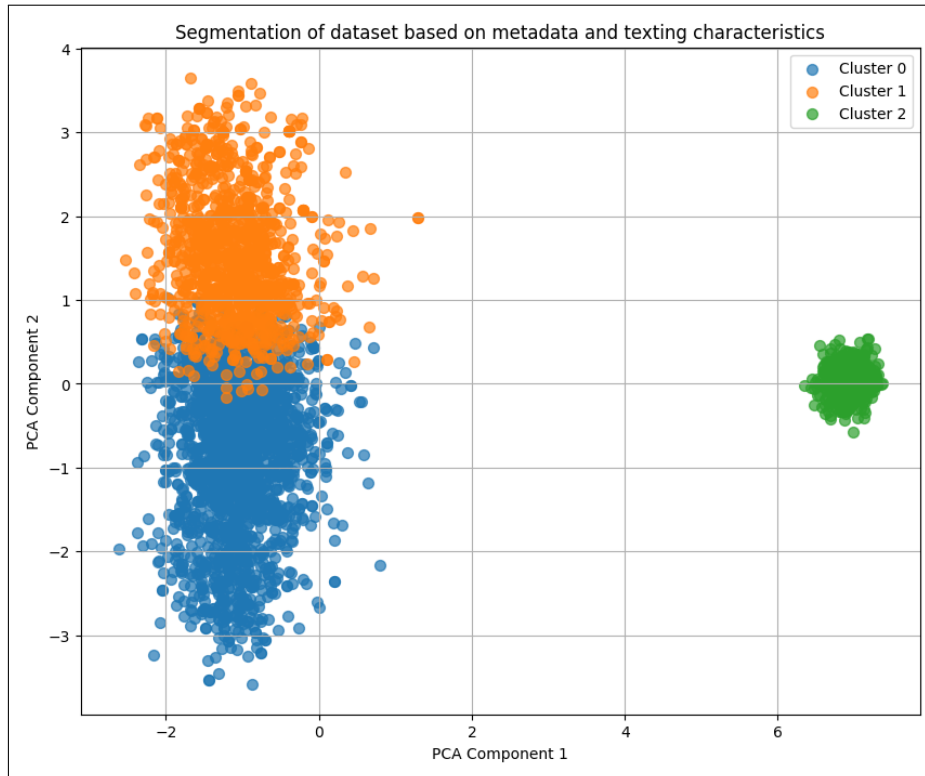
**Table 2.2:** Impact of gender, demography, age and profession on length of text messages (in characters).

		<i>Length of Text Messages</i>				
		Min	P 25	Median	P75	Max
Gender	Female	15.75	28.5	34.7	42	66
	Male	13	30.9	36.3	41.4	59
Demography	Urban	13	26.5	34.1	42.3	65.1
	Rural	15.75	33	36.6	41	66
Age	<=18	28.25	36.5	42.2	44.7	59
	19-40	13	29.7	35.8	41.4	65.1
	41-60	19.5	30.9	35.7	40.6	66
	>=61	24.8	26.5	30.6	30.6	33.7
Profession	Agriculture	15.7	33	37.6	41.25	55.3
	Service	15.8	25.2	31.4	37.1	60.9
	Shop Owner	13	30.1	37.5	46.75	66
	Student	32.6	41.3	43.2	44.6	59
	Unemployed	18.9	26.8	32.5	40.4	50.7

## 2.3 Modeling of Semi-Literate Personas: Analysis and Insights from SLT

We analyze survey responses by examining user characteristics (age, gender, education, language, profession, location, message frequency, etc.) and text message features (lexical, semantic, syntactic) to identify patterns, themes, and needs. The segmentation process is iterative and bidirectional: we first apply Machine Learning (ML) techniques, such as clustering, to group users, then validate and refine these clusters through user interviews. Feedback from these interviews is continuously integrated back into the ML models, creating a cyclical process until an equilibrium is reached where the segmentation is both statistically robust and practically grounded in user insights.

We used DBSCAN (with variations in  $\epsilon$  and *MinPts*) and K-Means (with k ranging from 1 to 5) as the clustering methods for this exercise. In order to identify the optimal number of clusters we leverage elbow method (plotting the within-cluster sum of squared errors for multiple values of k) as well as metrics like Silhouette Coefficient (comparing average Silhouette score across all data points for different values of k). We validate the identified clusters, with user interviews from each group along with feedback from our surveyors. After a few back and forth iterations, we baseline 3 user groups. The volumetric distribution of clusters (Cluster 0 - 55%; Cluster 1 - 31%; Cluster 2 - 14%;) along with pictorial representation (with reduced



**Figure 2.3:** Analysis of corpus on metadata and text message characteristics.

dimensions using PCA) is illustrated in Figure 2.3.

While one of the user group is peculiar and easy to isolate, the other two have some overlap and hence some of the models designed in our work may be common to both. We also conduct t-test as a statistical hypothesis test to determine if there is a significant difference between these groups. We use user characteristics (age, education, language) and text characteristics (euclidean normalization for semantic embeddings, readability scores and average length off messages) for t-test analysis. The results from this exercise are illustrated in Table 2.3 and it is evident that while Cluster 0 and 1 (Digitally Novice and Digitally Niche) differ statistically on 2/6 characteristics, the separation with Digitally Neglected happens on more parameters and hence is more distinct.

As a next step once the groups are finalized, for each cluster we carefully analyse the features and develop a detailed persona profiles that encapsulate the characteristics, needs, goals, challenges, behaviors, and preferences of the typical user. We provide each persona a name and include relevant demographic information to humanize them.

Once the personas are drafted we validate the same by cross-validating with target users, to ensure their accuracy and relevance. We leverage these personas extensively as a tool for decision-making in proposing AI-models, model design, and user



**Table 2.3: Statistical Differences Between Clusters: T-Test P-Values**

		Feature	t-statistic	p-value
Cluster 1	Cluster 0	<b>Age</b>	10.14725	0.0001
Cluster 1	Cluster 0	<b>Language</b>	16.33593	0.0001
Cluster 1	Cluster 0	Average Message Semantics	-0.78130	0.4356
Cluster 1	Cluster 0	Average Readability	-0.39122	0.6959
Cluster 1	Cluster 0	Average Length	-0.49830	0.6241
Cluster 1	Cluster 0	Education	0.03501	0.9721
Cluster 1	Cluster 2	<b>Average Message Semantics</b>	-5.14008	0.0001
Cluster 1	Cluster 2	<b>Average Length</b>	17.10925	0.0001
Cluster 1	Cluster 2	<b>Average Readability</b>	-7.44405	0.0001
Cluster 1	Cluster 2	<b>Age</b>	9.34453	0.0001
Cluster 1	Cluster 2	Language	-0.48131	0.6351
Cluster 1	Cluster 2	Education	-0.78130	0.4356
Cluster 0	Cluster 2	<b>Average Message Semantics</b>	-16.14926	0.0001
Cluster 0	Cluster 2	<b>Average Length</b>	21.97288	0.0001
Cluster 0	Cluster 2	<b>Average Readability</b>	-6.60472	0.0001
Cluster 0	Cluster 2	Language	-2.60284	0.0100
Cluster 0	Cluster 2	Age	-1.59245	0.1283
Cluster 0	Cluster 2	Education	-0.78130	0.4356

experience decision making.

### **Persona #1: Mr Rajiv Das**

#### **Demographics:**

Age - 62

Gender: Male

Occupation: Retired

Location: Suburban Pune, Maharashtra

#### **Background:**

Mr. Das is a retired individual with a background in unorganised hospitality services. He has limited exposure to technology throughout his career and has recently retired, finding himself in a world increasingly dependent on digital tools and devices. Mr. Das has attended school till class Ten skills but lacks confidence and familiarity with modern technology and English language.

#### **Goals and Objectives:**

- *Stay Connected:* Mr. Das wishes to stay connected with family and friends, especially since many of them now communicate through digital platforms.
- *Access Information:* He desires to access information online for news, hobbies, and health-related matters.
- *Manage Finances:* Mr. Das wants to learn how to manage his finances online, including online banking and bill payments.

#### **Challenges and Pain Points:**

- *Comprehensibility challenge:* Though Mr. Das reads and understands simple English, however he lacks confidence in complex words and often feels overwhelmed by the language used on internet
- *Limited Exposure:* Mr. Das has almost no exposure to internet technology throughout his life, making it challenging to adapt to the rapidly changing digital landscape.

### **Persona #2: Mrs Lakshmi Patil**

#### **Demographics:**

Age - 45

Gender: Female

Occupation: Agriculture

Location: Rural Konkan, Maharashtra

#### **Background:**

Ms. Patil is a farmer who has spent her entire life in a close-knit rural community. She is deeply connected to her native culture and language, which is the primary

means of communication within her community. Having never traveled outside her village, Ms. Patil has limited exposure to languages other than her own.

**Goals and Objectives:**

- *Agriculture Knowledge:* Ms. Patil seeks to share and receive agricultural knowledge within her community.
- *Community bonding:* She wishes to strengthen community bonds through storytelling, cultural events, and local festivities..
- *Access to Government Services:* Ms. Patil desires better access to government services and information available in her native language.

**Challenges and Pain Points:**

- *Language Barrier:* Ms. Patil faces challenges when browsing the government services that predominantly operate in regional or national languages.
- *Unengageable Digital Exposure:* She finds it challenging to use digital platforms due to the use of roman script used for digital content on her native language.

**Persona #3: Mr Raju Kumar****Demographics:**

Age - 43

Gender: Male

Occupation: Daily wage laborer

Location: Urban Pune, Maharashtra

**Background:**

Mr. Raju Kumar is a daily wage laborer who has had minimal formal education. He has basic numeracy skills along with ability to read and understand small words but struggles with reading at large. His exposure to digital devices is limited, and he faces considerable challenges in using them due to his low literacy level.

**Goals and Objectives:**

- *Communication:* He desires a simple way to communicate with family members and receive updates from his family living in the village.

**Challenges and Pain Points:**

- *Low Literacy:* Mr. Kumar's literacy level is minimal, making it challenging for him to read written information or navigate digital interfaces.
- *Digital Intimidation:* He feels intimidated by the complexity of digital devices.

## 2.4 Persona Analysis: Defining Strategic Objectives for our work

As the next step, we brainstorm, explore and shortlist the potential solutions which can address the challenges defined for each persona and fulfil their objectives, and achieve research maxims and other considerations illustrated in Chapter 1. We validate the shortlisted ideas to ensure its potential success and effectiveness, on three pillars of ideation illustrated below to make informed decisions.

- **Desirability** - Validate if the idea suggested meets the needs and preferences of the target audience and we are solving the core problem faced by end users.
- **Viability** - Validate if the solutions are sustainable and extendable along with long term (futuristic) practical usefulness of the value chain.
- **Feasibility** - Validate if the idea proposed is technologically and operationally doable.

We shortlist three ideas for our work, which forms the basis of rest of the thesis. All of these are explored individually and we have published work on each one independently. All of these address the sub-problems defined.

- For persona #1, referenced as *Digitally-Novice* in our work, we propose a paraphrasing system for semantic and morphological simplification, using transformer architecture along with late-fusion algorithm for the application of multiple pivots.
- For persona #2, referenced as *Digitally-Niche* in our work, we propose multiple AI-models for digital on-boarding of low (and ultra-low) resource languages and scripts along with generalised methods to create large volume, good quality bilingual dataset creation, and create foundational ecosystem and enable higher tasks.
- For persona #3, referenced as *Digitally-Neglected* in our work we propose design of a multimodal visual communication framework and working prototype enabling non orthographic methods of communication using Generative AI and Large Language Models (LLMs).

Since there is an overlap between *Digitally-Novice* and *Digitally-Niche* user groups, so it is possible that the models suggested for digitally-niche communities may also help some digitally-novice users.

## **2.5 Ethics Statement**

Participation in the survey has been voluntary. Informed consent in writing has been obtained from all survey participants. The privacy rights of human subjects has been observed. No personal identifier information (PII) or clinical data has been collected as part of our survey. The research follows the “Guidelines for Ethical Considerations in Social Research & Evaluation in India” as described by CMS [62]. A self-administered Ethics Sensitivity Test was conducted based on the guidelines and grade of “Very Good” has been observed. From the best of our understanding there is no additional approval needed on ethics to use this data for research purpose.



## **Part 2**

# **Enabling "Digitally-Novice" Semi Literates using Semantic Simplification**





# 3 | Multi-Pivot Sequence-to-Sequence Transformer with Late Fusion Architecture for Reductive Paraphrasing

*“The ability to simplify means to eliminate the unnecessary so that the necessary may speak.”*

---

Hans Hofmann

This chapter presents the design, implementation, and validation of a novel AI model for text simplification in the context of semi-literates, aimed at making digital information more accessible to them [25]. The model incorporates advanced transformer techniques (multi-pivot ensemble and late fusion architecture) tailored to the linguistic and cognitive needs of semi-literate users, ensuring that the simplified text uses the vocabulary relevant to them and alongside retains its original meaning and clarity.<sup>1</sup>

## 3.1 Introduction

In today’s increasingly digitized world, access to information and communication plays a pivotal role in fostering social inclusion and equitable participation. However, a significant portion of the population faces barriers when encountering complex language structures, hindering their ability to comprehend and engage with textual online content effectively. This challenge is particularly pronounced among individuals with lower academic literacy, who had restricted schooling. We refer this group of

---

<sup>1</sup>Parts of this chapter have been published in the Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (FIRE), ACM digital library

semi-literates as *Digitally-Novice* audience and design the necessary simplification AI-models to make digital content engageable and accessible for them. It is noteworthy that the proposed models can also help individuals with cognitive or learning disabilities due to similar challenges faced.

The design of content simplification goes beyond surface level lexical simplification and finds a sweet spot, interdisciplinary collaborative research between linguistic theories and Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP) algorithms. While the role of AI technology in fostering digital accessibility is primary however understanding the universal linguistic complexities and reductive paraphrasing via linguistic studies (theories and principles) shall be equally relevant. Bringing the two together we propose a reductive semantic algorithm, empowering individuals with lower academic literacy levels to navigate the complexities of the digital age.

### 3.1.1 Defining Research Objectives: AI-Models for Enhanced Semantic Cohesion in Sentence Simplification

The goals or overarching objectives for this work is focused on developing AI-models to paraphrase complex English into simplified English for individuals with lower literacy levels, enhancing accessibility and fostering digital inclusion. We define clear objectives for pointed focus and direction, and validate the proposed models towards the completion on these objectives.

- **Reduced semantic complexity:** In order to enhance digital communication and interaction between individuals with lower literacy levels, it is a must to bridge the gap between the existing complex content on digital platforms with the reduced vocabulary understandable by digitally-novice participants. It is also important to identify appropriate metrics and empirically validate the reduction in semantic complexity using the proposed models.
- **Preserving Meaning and Clarity:** It is to be noted that there should be minimised loss of information while paraphrasing. AI models and algorithms proposed should prioritize preserving the essential meaning and clarity of the original text during the simplification process. This includes identifying and retaining key information, concepts, and nuances while removing unnecessary complexity and ambiguity.

In a nutshell, the proposed models should achieve semantic simplification of complex English into simple sentences and facilitate an inclusive digital environment for digitally-novice semi-literates, reduce frustration and stay engaged with the digital content. The targeted users already have enough motivation and know-how of the digital forms of information, but the semantic complexities keep them disengaged and disinterested.

## 3.2 Assessing State-of-the-Art Research Trends and Identifying Gaps in Reductive Paraphrasing Methods

While extensive research has been conducted on paraphrasing, utilizing various methodologies such as lexicon-based or structural approaches, and employing different types of data sources including pivot-based (zero-shot) or parallel corpora, there has been limited investigation into reductive paraphrasing. Reductive paraphrasing involves design of simplified vocabulary, and build paraphrasing systems in order to reduce complexity of sentences.

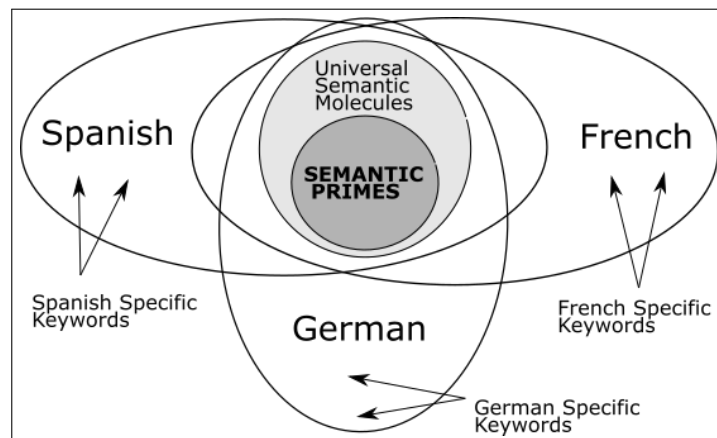
### 3.2.1 Review of SOTA Computational and Linguistic Methods

In our work, we leverage both computational and linguistic methods in conjunction, using linguistic theories to build the theoretical foundation of semantic simplification and the use of computational linguistics methods including deep learning based transformer models, to execute the proposed design.

#### **Linguistic methods**

##### **Natural Semantic Metalanguage (NSM):**

The linguistic theory of NSM argues that all human languages share a common set of simple and global semantic concepts, referred as "semantic universals". These semantic universals are innate and culturally universal, serving as the foundational elements for expressing meaning in language. By combining these basic entities, speakers can construct complex meanings and communicate effectively across different languages and cultures. NSM is considered as the most comprehensive, practical and acceptable linguistic theories for reductive semantics [19, 20].



**Figure 3.1:** Schematic representation of relationship between three languages as outlined in the linguistic theory of NSM [19]

NSM further classifies semantic universals into (a) semantic primes which form the core foundational concepts, and can be considered as the elements in the periodic table for core concepts; (b) semantic molecules, which are complex concepts formed with primes and may have different properties than its constituents; (c) semantic templates, which are universal category of semantic concepts and inherently are composed of same/similar semantic primes/molecules. We illustrate this in Figure 3.1 for reference.

#### **Other linguistic theories:**

There are additional linguistic theories on semantic simplification including the theory on Plain English [63], Universal Design for Learning (UDL) [64], Controlled English [65], Simplified Technical English (STE) [66] etc. However most of these are domain specific and may not be suitable for universal use, specially in the context of digitally-novice semi-literates.

## **Computational methods**

While there are multiple methods for reductive paraphrasing and simplification of text using a variety of computational linguistics techniques, like modifications on surface lexical forms, morphology based simplification, syntax based simplifications and discourse based changes (as illustrated in Chapter 1), our work leverages the principles of Zero Shot Learning (ZSL), Pivoting and Neural Machine Translation (NMT) using late fusion transformer architecture for decoder convergence.

#### **Zero-shot Learning (ZSL):**

ZSL represents a learning framework utilized in scenarios where the available training and test datasets are disjoint. This paradigm capitalizes on auxiliary information to train models due to the absence of directly labeled training data. In recent years,

ZSL has emerged as a crucial facilitator in machine learning, particularly in the context of training deep learning models that require substantial volumes of training data. It has been observed that ZSL can play an important role in machine translation and paraphrasing tasks [67, 68, 69].

### **Pivoting:**

Pivoting stands as a widely adopted technique in the implementation of ZSL, particularly in addressing the data challenges posed by the deep learning methods. This method helps in resolving translation ambiguities significantly by leveraging translations into a third language, known as a pivot. This pivotal observation paved the way for the incorporation of pivoting techniques in various language processing tasks, including reductive paraphrasing. Subsequent research endeavors have further elucidated the efficacy of multi-way pivoting strategies, in delivering notable performance enhancements over traditional single-pair pivoting models [70, 71, 72, 73, 74].

### **Neural Machine Translation (NMT):**

NMT is a very popular deep learning technique for translation/paraphrasing tasks, and runs on the encoder-decoder architecture. The encoder transforms the source text into continuous space representations, which are then utilized by the decoder to produce the target sentence. To augment this method, an attention mechanism is employed, focusing on specific regions for improved performance. NMT leverages the advantages such as reduced memory usage and streamlined decoding processes, facilitated by techniques like beam search.

Illustrating with an example, for a multi-pivot, English paraphrasing task, consider the input sentence ( $Eng_{in}$ ) using the foreign pivot languages set ( $\Gamma$ ) to finally convert to paraphrased sentence ( $Eng_{para}$ ), the model can be mathematically represented as,

$$\Gamma = \{Pivot_1, Pivot_2, Pivot_3, \dots, Pivot_n\} \quad (3.1)$$

$$P(Eng_{para}|Eng_{in}) = P(Eng_{para}|Eng_{in}, \Gamma) = P(Eng_{para}|\Gamma) \quad (3.2)$$

In the normal case of using a single encoder and a decoder along with attention mechanism, the input sentence is encoded into to a context vector which is later decoded to target language taking into consideration the attention mechanism. However for the use of multi-pivot scenario, multiple encoders work together to convert input sentence into multiple context vectors. Each context vector is fed into a separate decoder. However all the decoders converge to single output sentence using two conversing techniques (i) Early averaging and (ii) Late averaging.

*Early averaging* is to perform averaging of the multiple translation paths when computing the time dependent context vector. At each time  $t$  in the decoder, the time-dependent context vector for each source language is computed as:

$$C_t = \lambda_1.C_{t1}^{P_1} + \lambda_2.C_{t2}^{P_2} + \lambda_3.C_{t3}^{P_3} \dots \lambda_n.C_{tn}^{P_n} \quad (3.3)$$

$C_t$  is the final computed context vector,  $C_{ti}$  is the individual context vector for each pivot language  $P_i$  and  $\lambda_i$  are the weights of multiple languages summing to 1. The decoder's hidden state also needs to be the average of the initializes of the n encoders.

*Late Averaging* on the other hand is to merge multiple translation paths by taking average of the probability distributions of tokens, and using softmax in the output layer.

$$\begin{aligned} p(y_t = w|y < t, \Gamma) = \{ & \lambda_1.p(y_t = w|y < t, P_1) + \\ & \lambda_2.p(y_t = w|y < t, P_2) + \\ & \lambda_3.p(y_t = w|y < t, P_3) \dots + \\ & \lambda_n.p(y_t = w|y < t, P_n) \} \end{aligned} \quad (3.4)$$

It has been observed that late averaging works better specially in case of paraphrasing scenarios [75, 73].

### 3.3 Critical Analysis: Unaddressed Issues and Opportunities in Reductive Semantics

While there baseline methods to achieve text simplification and the use of deep learning based transformers for sequence to sequence tasks like traslation and paraphrasing, however the contextual problem of semi-literates (digitally-naive) who need simplified vocabulary has not been directly addressed. The primary issues with existing methods is described below:

- **Non availability of semi-literate contextual vocabulary:** Having a large size vocabulary corpus that is understandable by people with less education can assist with training Machine Learning (ML) models, scalability, diverse representation and a valuable resources for bench-marking. It is very pivotal to have a set of sentences in easy to understand language for building the enabling software for digital enablement of digitally novice semi-literates.
- **Interdisciplinary research combining linguistics and AI models for semi-literates:** While the study on linguistics help with cultural and contextual

factors on how humans comprehend and process language, the AI ecosystem (including ML and NLP algorithms) gives computational leverage to harness large volumes of data and information. Bringing the two together, the complex challenges of text simplification for digitally-novice users can be addressed, ultimately enhancing accessibility to information and promoting digital inclusion for all.

### 3.4 Novel Zero-shot Reductive Paraphrasing Methodology using Multi-Pivot late fusion Transformer Model

We propose a novel zero-shot learning approach to achieve reductive paraphrasing for semi-literates, using late averaging transformer architecture. The experiment results demonstrate that the proposed method helps to achieve 25% reduction in education level comprehensibility.

#### 3.4.1 Semantic Ontology-Based Extraction and Construction of a Large-Scale Training Dataset

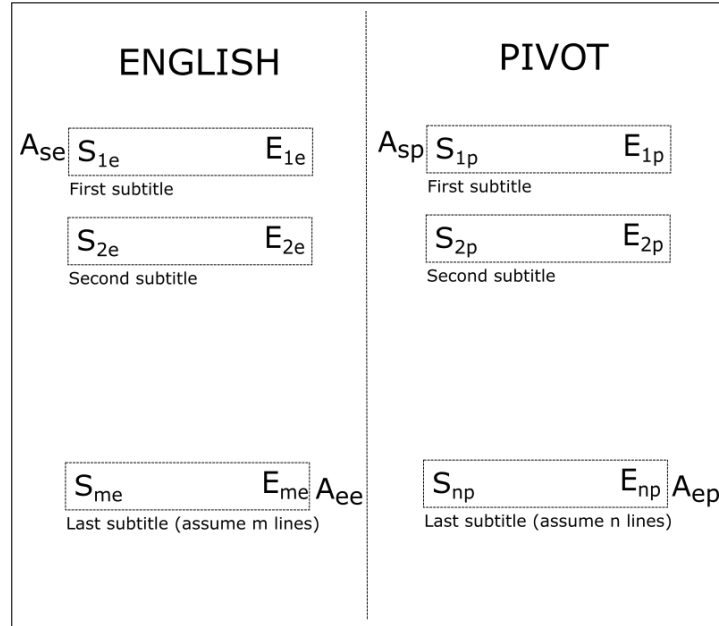
A wealth of paraphrasing data is readily accessible from diverse sources and meticulously aligned across multiple languages. However, the majority of available datasets lack categorization based on reductive semantic simplicity. We work towards building a dataset for individuals with limited literacy skills. Our dataset construction process entailed two primary phases: (a) the compilation of a simplified vocabulary and (b) the assembly of the dataset corpus. Furthermore, we curated a distinct dataset specifically designed to empirically evaluate our model's performance and ascertain its suitability for users with limited literacy proficiency.

#### Building the Ontology

The philosophy we leverage to design and build our dataset can be summarised as, *"The conversation in the animation movies for younger audiences, which is released globally across geographies is carefully thought through to keep the exposure to vocabulary minimal. These dialogues can be leveraged to create initial vocabulary for the needs of digitally-novice communities and facilitate their digital empowerment."*

OpenSubtitles is one of the largest collaborative subtitle databases on the internet managed and maintained by volunteers and contributors. It provides free access to

a vast collection of movie and television show subtitles in multiple languages [76]. We leverage this platform and download the subtitles for a small set of animation movies<sup>2</sup> in four languages (English, French, German and French).



**Figure 3.2:** Sentence alignment methodology across subtitle files of two languages.[77]

In most cases, not all subtitle files (for the same movie, across various languages) have the same frame rates and hence the dialogues cannot be mapped directly. Some amount of pre-processing including calculation of offsets, and realignment of subtitles on a single time frame is required to be done in order to map the dialogues correctly. Figure 3.2 illustrates the methodology used for sentence alignment across English and Pivot language combinations, which is further elaborated in equations 3.5/3.6. We use subtitle start ( $A_{se}$  and  $A_{sp}$ ) and subtitle end ( $A_{ee}$  and  $A_{ep}$ ) as anchor points, and calculate offset ( $\Delta$ ) to fine tune the subtitles.

$$Offset = \Delta = \frac{A_{ep} - A_{sp}}{A_{ee} - A_{se}} \quad (3.5)$$

For each subtitle in each file at position  $i$  calculate lag ( $\lambda$ ), based on the difference of end position  $E_{ie}$  and start position  $S_{ie}$

$$Lag = \lambda = E_{ie} - S_{ie}$$

$$S_{ie} = [(S_{ie} - A_{se}) * \Delta] + A_{sp}$$

<sup>2</sup>Big Hero, Boss Baby, Captain America, Coco, Dora and the Lost city of Gold, Finding Nemo, Frozen, How to train your Dragon, Lion King, Minions, The Lego Movie, The Lego Movie 2: The Second Part, Toy Story



---

**Algorithm 1:** Our adaption of sentence mapping[77]

---

**Result:** Mapped sentence pairs

```

1 initialize: mapped_sentences ← null;
2 while source_file does not reach EOF do
3   | line_src ← nextLine(source_file);
4   | sliding_window_start ← end_time(line_src − 1);
5   | slidng_window_end ← start_time(line_src + 1);
6   | for target_line ≥ sliding_window_start; target_line ≤ =
   |   | slidng_window_end; do
7   |   | mapped_sentences ← line_src, target_line
8   | end
9 end

```

---

$$E_{ie} = S_{ie} + \lambda \quad (3.6)$$

We observe that using the method as proposed, and using random sampling to validate the effectiveness the mapping becomes 90% accurate. Our final repository of simplified ontology, contains more than 10,000 tokens in English language. The tokens for other pivot languages (German, Spanish, French) also are in the same range. We also illustrate the entire method as Algorithm 1, to illustrate the entire process in detail and reduce ambiguity.

## Building the Test/Train Datasets

Deep learning methods require large volume of data for training purpose. While the initial ontology works like a baseline for restrictive vocabulary, it is required to have large volume of Pivot-English language sentence pairs for training purpose. The Open Parallel Corpus (Opus) maintained by University of Helsinki and the Language Bank of Finland is a large collection of parallel texts in multiple languages. It contains translations of various types of text (literature, news articles, etc.) aligned in parallel to facilitate multilingual research and language-related tasks [78]. We leverage Opus corpus, for extracting sentence pairs of Pivot-English and apply two filters (i) Use only those sentences where all words are within the initial vocabulary of interest; (ii) discard longer sentences, since semi-literates refrain from using longer sentences (inference from the survey, as illustrated in Chapter 2). Our final training corpus, as illustrated in Table 3.1 contains more than 250K sentence pairs which are sufficient to train the decoders in our model.

It is also important to have a test dataset to empirically validate the effectiveness of our work. To the best of our knowledge there is no ready to use dataset contextualised for semi-literates. While "Turk corpus" [79] can become the nearest available option, but it also needs to be fine tuned and contextualised for this purpose. For this

**Table 3.1: Training dataset**

Language combination	Sentences	Tokens (English)
French-English	265,328	9,931
German-English	272,858	10,095
Spanish-English	268,409	9,802

work, we employed a group of five human annotators (classified as digitally-naive semi-literates) from urban and rural regions in Maharashtra, India. We also used a subset of sentences (filtered by size and context) from Turk Corpus, to enrich our test dataset.

### 3.4.2 Design of Multi-Pivot Zero-Shot Reductive Semantic Paraphraser

As illustrated in Figure 3.3, the design of the paraphraser, follows a three step process (i) Translation to Pivots, (ii) Lexical Simplification using Complex Word Identification (CWI), and (iii) Morphological simplification via late-averaging decoders.

#### Translation to Pivots

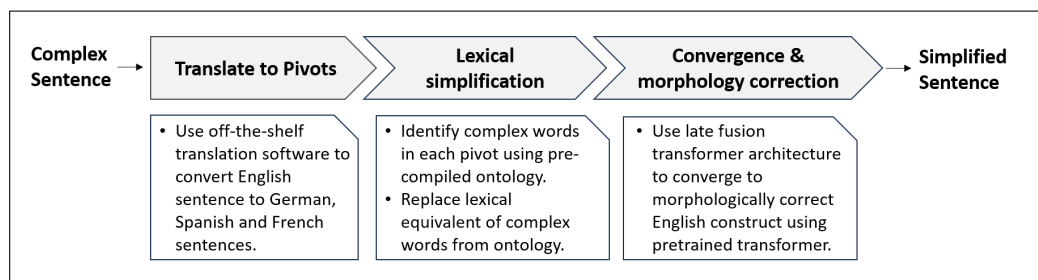
Our methodology incorporates multi-pivot translation and convergence, commencing with the translation of the complex input sentence into its respective equivalents in German, Spanish, and French. To accomplish this initial step, we employ the widely used "Google Translate" tool. To further illustrate and clarify the operational framework of our system we use a test example and run it through various segments of the pipeline.

Input sentence - "The atmospheric conditions are magnificent EOL"

Translated sentences(s) to multi-pivots:

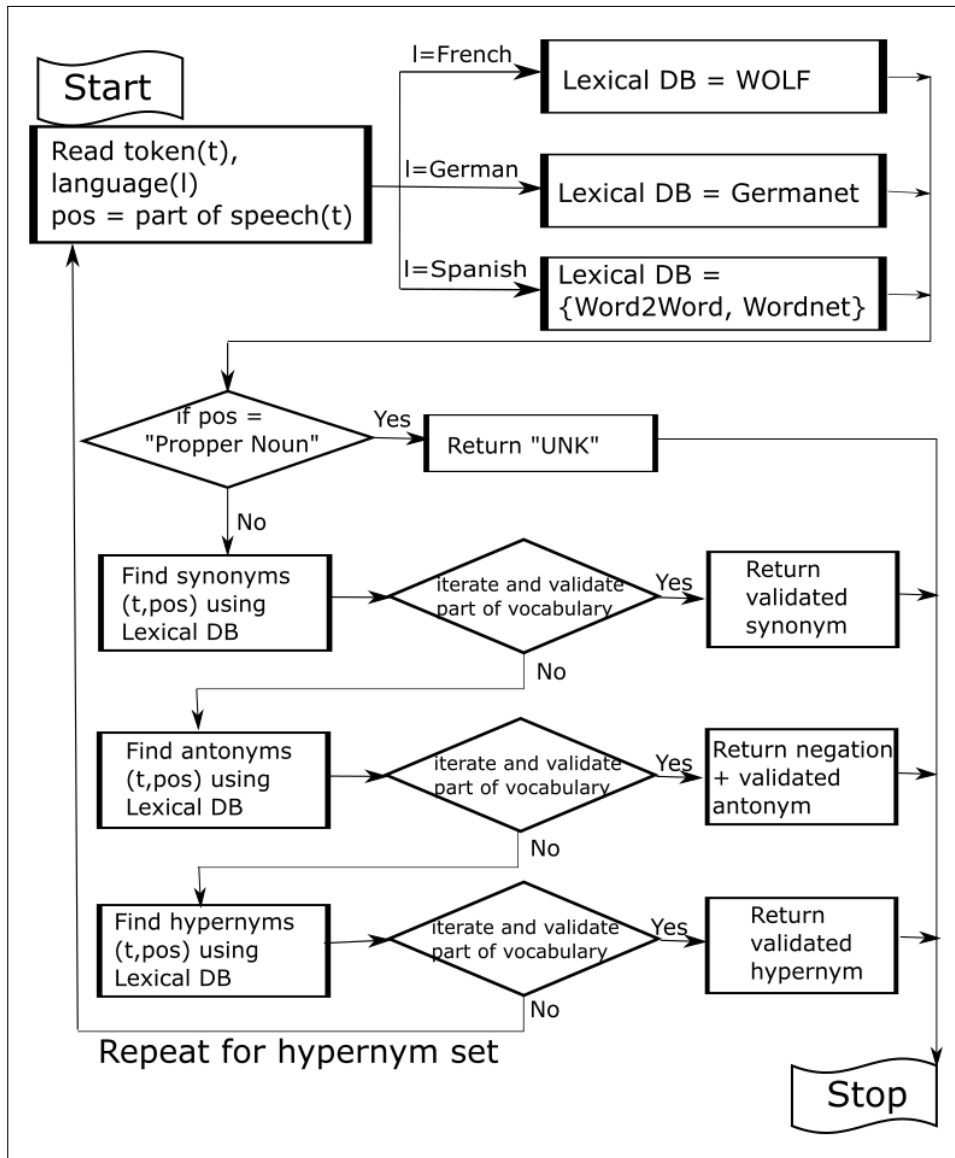
De- *"Die atmosphärischen bedingungen sind großartig EOL"*

Es- *"Las condiciones atmosféricas son magníficas EOL"*

**Figure 3.3: High level flow diagram of Reductive Semantic Paraphraser**

Fr- "Les conditions atmosphériques sont magnifiques EOL"

## Lexical Simplification



**Figure 3.4:** Lexical Simplification procedure workflow

For each translated sentence, it is important to identify the complex words and replace the complex words by semantically simpler words and achieve lexical simplification. This process can be broken down into two sub parts namely (a) Identification of complex words and (b) Lexical substitution.

Complex Word Identification (CWI) represents a pivotal stage in our framework. It's imperative to identify complex words within a sentence before embarking on

lexical simplification for each. In our experiment, we validate the translated sentences from the previous step with our baselined vocabulary, and tokens not found are subsequently marked as complex.

Following the same example, as illustrated in previous section, we mark complex words for lexical simplification.

De- "*Die atmosphärischen bedingungen sind großartig EOL*"

Es- "*Las condiciones atmosféricas son magníficas EOL*"

Fr- "*Les conditions atmosphériques sont magnifiques EOL*"

For each complex word identified, lexical simplification must be conducted individually for each pivot language. This entails simplifying all complex words identified in the previous stage to terms within our designated vocabulary. We leverage a variety of lexical databases, that organizes words into sets of synonyms, antonyms, and other word senses (hypernyms, hyponyms, meronyms etc.) and describes the semantic relationships between them. Lexical databases are generally available for popular languages and popular in NLP tasks (we use WOLF [80] for French, GermaNet [81] for German, and a combination of Word2Word [82] and Wordnet [83] for Spanish). We illustrate this method as a flow chart in Figure 3.4.

Following the example, the lexically simplified sentences in pivot languages becomes.

Fr- "*Les conditions météo sont magnifiques EOL*"

De- "*Die bewegungslos Bedingungen sind großartig EOL*"

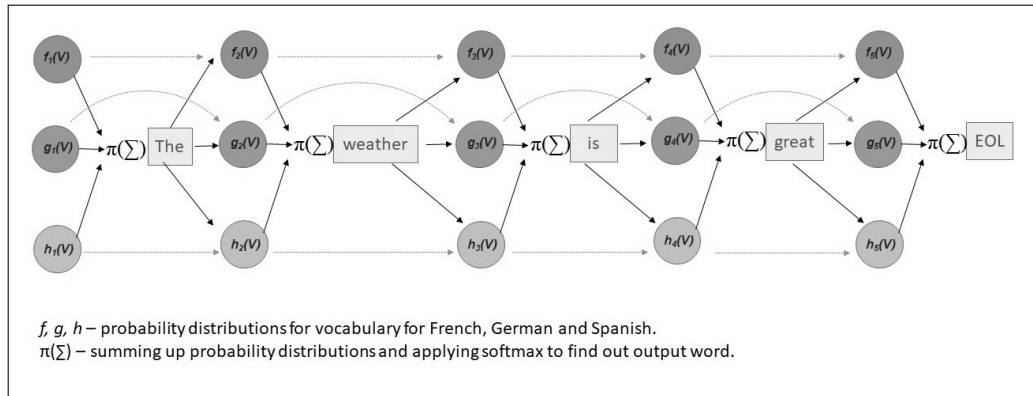
Es- "*Las condiciones clima son magníficas EOL*"

## Convergence and morphological correction

This marks the final stage in our design for reductive semantics, and helps to converge independently simplified sentences to morphologically correct single English sentence form. While the input sentence splits into multiple pivots and surface level lexical simplification happens with replacement method, it is likely that some morphological/syntactic distortions may happen. Leveraging the principles of the use of late fusion transformer, as mathematically illustrated in Section 3.2.1 we converge three independent decoders into a single final output. Further illustrated and graphically represented in Figure 3.5 the probability distributions from each decoder is combined in the final layer using soft-max function. While the individual decoders (for each Pivot-English language pair) are trained on training dataset (Table 3.1), late-averaging helps to bring these decoders together to construct the final output.

Further processing the illustrated example, the end result of our semantic paraphraser becomes,

*"The weather is great EOL"*



**Figure 3.5:** NMT using Late Averaging and 3 languages pivot model

We illustrate further examples from our work in Table 3.2 to indicate the reduction of semantic complexity and effectiveness.

### 3.5 Evaluation of proposed AI Model Performance on Effective Sentence Simplification and Preserving Semantic Integrity

Empirical testing for reductive semantics needs to be clearly discussed, defined and appropriate metrics need to be classified to measure the same. It is also advised to have some bench-marking done with SOTA methods to compare the results with industry standards. We use three parameters to measure the effectiveness of our work.

- **Simplicity:** Since the intent of our work is digital enablement for semi-literates, it is very pivotal to measure the simplicity of paraphrased sentences.
- **Adequacy:** It is important to make sure that the paraphrases sentences do not alter the semantics of the sentence.
- **Fluency:** It is also important to make sure that the syntax and morphology of the paraphrased sentences is intact, preferably also simplified.

For the choice of metrics, we use BLEU (Bilingual Evaluation Understudy) and SARI (Semantic Adequacy, Relevance, and Information) metrics is to evaluate the quality of machine-generated text [84, 85]. These metrics provide objective and quantitative measures of translation and simplification quality, in terms of adequacy and fluency. For measuring simplicity of paraphrased sentences, we use a variety of

**Table 3.2:** Illustrative examples of complex and converted paraphrased sentences from our work

Original Sentences	Paraphrased sentences
Some medications are far too much for impoverished communities.	Some drugs are too much for poor groups.
Religious movements periodically declared the carols as sinful.	Religious groups consider Christmas songs as sin.
There is no facile solution.	No simple solution.
A certain austere simplicity was noticeable.	Some simplicity was visible.
His own share in the proceeds was about a hundred thousand dollars.	Their own participation in profits was about one hundred thousand dollars.
Long before writing and books were in common use, proverbs were the principal means of imparting instruction.	Long before writing and books were normally used, popular sayings were means of transmitting the instruction.
The officials expressed concern about reigniting longstanding Mexican concerns.	The officials expressed concerns about influence of Mexico.
One faction declared it would begin an armed struggle.	A pack said it will begin fight arms.
Therefore he made up his mind to accede to his uncle's desire.	So he decided it to agree the wishes of the uncle.
The prejudice society is always against free thinkers.	The company is always against free mind.
Vacation homes typically have pristine sand and crystalline waters.	Rest houses have impeccable sand and clear waters.
He settled in London, devoting himself chiefly to practical teaching.	He lived in London involved mostly in teaching.
He left a detachment of 11,000 troops to garrison the newly conquered region.	He left 11,000 people to make strong new location.
We went to some cheesy bar in Pune.	We went to a cheap hotel in Pune.

readability scores to assess the readability or ease of comprehension of paraphrased sentences.

While BLEU is preferred for evaluating machine translation and general paraphrasing tasks where text simplification is not a primary consideration, SARI is a more appropriate metric for assessing lexical simplicity by comparing the model's output to multiple simplification references. We use SARI for our evaluation [86].

SARI measures the goodness of three types of operations, namely (a) Additions: Words that are added to the simplified sentence; (b) Deletions: Words that are removed from the original sentence; (c) Keeps: Words that are retained from the original sentence. For each of the three operations (Additions, Deletions, Keeps), SARI calculates an F1 score (or precision and recall in some variations) and then averages

these scores. The overall SARI score is the average of the scores for these operations.

$$\text{SARI} = \frac{\text{Add}_F + \text{Keep}_F + \text{Delete}_F}{3}$$

where each F1 score ( $\text{Add}_F$ ,  $\text{Keep}_F$ ,  $\text{Delete}_F$ ) is calculated as:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table 3.3 illustrates multiple paraphrasing models tested on "Turk Corpus" ranked by SARI score [87, 88, 89, 90, 91, 92]. None of these models have been designed for reductive semantics to be used for digital enablement for semi-literates, however are the closest reference for bench-marking purpose. We can observe a SARI score between 33 to 42 in the illustrated reference. Our proposed model gives the SARI score of 36.65 which is within the SOTA range and works effectively for our targeted users.

While SARI measures the adequacy and fluency of our work, we use readability scores to measure simplicity of paraphrased sentences. Simplicity can give an indication of accessibility for those with low literacy levels, cognitive disabilities, or English as a second language. It is also observed that texts with higher readability scores are generally easier to comprehend, leading to better information retention, reduced misunderstandings, and improved overall communication outcome.

For our evaluation, we use a variety of readability scores including, (i) "Flesch Reading Ease" evaluates the ease of reading on a 0-100 scale and a score of more than 70 qualifies the text to be easy to understand by most people with limited vocabulary; (ii) Readability score referred as "Gunning Fog Index" is another alternative to measure readability and score estimates the years of formal education a person needs, to understand the text on the first reading; (iii, iv and v) "Flesch Kincaid Readability", "Dale Chall Readability Score" and "Coleman Liau Index" scores on estimating

**Table 3.3:** SARI scores for various state of the art models

Model	SARI
ACCESS	41.87
DMASS + DCSS	40.45
Edit-Unsup-TS	37.85
DRESS	37.08
NSELSTM-S	36.88
SEMoses	36.7
<b>Our Work</b>	<b>36.65</b>
NSELSTM-B	33.43

**Table 3.4:** Impact of paraphrasing on readability indexes.

Reading Index	Original	Paraphrased	Preferred Value	Improvement
Flesch Reading Ease	55.72	73.49	Higher	31.89%
Gunning Fog Index	9.79	7.55	Lower	22.88%
Dale Chall Score	8.70	7.26	Lower	16.55%
Flesch Kincaid Grade	8.17	5.41	Lower	33.78%
Coleman Liau Index	10.70	8.32	Lower	22.24%
<b>Average Improvement</b>				<b>25.47%</b>

grade level for sentence comprehension. The purpose to use a variety of these metrics is to provide a more comprehensive and robust evaluation of our system's performance. We illustrate the results from this exercise in Table 3.4 and observe more than 25% average improvement in comprehensibility, which is a major improvement from initial sentences. The mathematical interpretation of these readability scores are illustrated in Appendix B.

## Ablation studies

In order to substantiate the effectiveness and contribution of late fusion transformer approach using late-averaging method in our paraphraser we conduct ablation studies using single pivot independently, and comparing the same with multi-pivot approach. As illustrated in Table 3.5, it is evident that having multiple pivots is instrumental in overall effectiveness of our work.

**Table 3.5:** Comparing single Vs multi-pivot. ARII (Average Reading Index Improvement) is the average of all readability scores, to measure the simplicity of paraphrased sentences.

Pivot Language	SARI	ARII
Only French(Fr)	35.29	17.61%
Only German(De)	33.07	20.74%
Only Spanish(Es)	35.83	18.77%
All 3 (Fr, De, Es)	36.67	25.47%

In the end we perform random sampling through human evaluation to validate the principles of linguistic theory of NSM, checking for "Language specific keywords" in input and output sentences. It was interesting to observe that language agnostic concepts are maintained in the final paraphrased sentences and specific words are eliminated. As an example the input sentence "We went to some cheesy bar in Pune" becomes "We went to a cheap hotel in Pune", and the keyword *cheesy*, which is popularly used in "English speaking native context" gets replaced with a more universal keyword "cheap" which is understandable universally.



## 3.6 Conclusion and Limitations

The proposed model as described in this chapter, introduces an innovative reductive semantic paraphrasing pipeline using zero-shot learning paradigm aimed at facilitating digital enablement for digitally-naive semi-literates. The experimental findings illustrate that our proposed approach leads to a significant one-quarter reduction in education levels required to comprehend the initial content effectively. Furthermore, we validate that employing late-fusion transformer architecture, yields superior results compared to using a single pivot language alone.

We acknowledge the inherent subjectivity involved in crafting simplified vocabulary tailored for semi-literate individuals within the Indian context. While leveraging universal simple vocabulary derived from movies targeted at younger audiences presents a promising and scalable approach for constructing a simplification system, it is essential to address potential disparities that may arise between this universal lexicon and the linguistic nuances specific to semi-literate Indian-English speakers. As part of our forthcoming research endeavors, we aim to enhance the precision of our corpus by incorporating crowd-sourced input and conducting a comprehensive re-evaluation of our findings to uphold their academic rigor and validity.



## **Part 3**

# **Enabling "Digitally-Niche" Semi Literates through Automatic Data Generation and Augmentation**



# 4 | Versatile Low Resource OCR Methodology using Contrastive Learning, GAN Augmentation, and Auto Glyph Feature Extraction

*“Every language is an old-growth forest of the mind, a watershed of thought, an ecosystem of spiritual possibilities.”*

---

Noam Chomsky

In this chapter, we delve into a novel and versatile methodology of Optical Character Recognition (OCR) we have developed for remote scripts [93]. Remote scripts pose unique challenges due to their distinct characters and writing systems, often requiring specialized techniques for accurate text extraction and recognition. By designing OCR capabilities for remote scripts, we aim to bring more diversity to digital content, promote language preservation, and support linguistic research in marginalised semi-literate digitally-niche communities.<sup>1</sup>

## 4.1 Introduction

Digital inclusion for people who use languages other than those predominantly featured on digital platforms is essential for fostering diversity, equity, and accessibility in the digital realm. Language serves as a fundamental aspect of identity and communication for individuals and communities worldwide and it not appropriate to expect that every community should learn the languages of the internet. We refer the the users of minority or less commonly represented languages as "*Digitally-Niche*"

---

<sup>1</sup>Parts of this chapter have been published in the proceedings of EACL 2024 [93]

semi-literates who in spite of being academically educated in their native language face barriers in accessing digital platforms. Enabling and on-boarding marginalized languages and scripts in the digital ecosystem can help digitally-niche users with access to information, participating in online discourse, and utilizing digital services effectively.

From the socio-linguistic point of view, digital inclusion is imperative for every language, regardless of its geographical or cultural origins, as it represents a unique worldview and cultural heritage. Digital inclusion facilitates the preservation, promotion, and revitalization of linguistic diversity, fostering cultural identity and social cohesion.

Architecting AI-models to enable low resource languages and scripts is essential for upholding linguistic rights and empowering "*Digitally Niche*" audience. In this chapter we illustrate the design of a novel methodology to architect OCR pipeline for scripts where no (or minimal) digital data is available.

### 4.1.1 Defining Research Objectives: Towards Establishing Robust OCR AI-Models with Limited-to-None Digital Training Corpus

While the primary objective of our work is to build AI technology to enable digitally niche semi-literates, however we define more granular objectives of this work to focus on key outcomes from this.

- **Digital on-boarding for scripts with no electronic footprint:** Build an universal methodology to enable digital on-boarding for remote scripts (and languages which use these scripts). This will help in an inclusive and responsive needs of digitally-niche communities leading to improved access to education, healthcare, and economic opportunities.
- **Building sizeable good quality datasets for downstream AI applications:** Preservation of linguistic diversity and cultural revitalization is only possible if researchers work together and build re-usable assets. In today's world, the cornerstone of advancing AI research in language lies in leveraging large, high-quality datasets to train complex models and ensure robust performance. The objective of our work, is also to build good quality large digital dataset methodology and enable higher computational linguistics tasks.

With on-boarding and enablement of remote scripts on digital platforms, digitally-niche communities can use digital devices for faster and effective communication

within their communities. It shall facilitate access to information, services, and opportunities, empowering individuals and communities to participate more fully in social, economic, and political life.

## 4.2 State-of-the-Art in OCR: Literature Review and Gaps for Low-Resource Script Digitization

Digital on-boarding for a new script or language starts with the foundational task of converting handwritten or printed archives (non digital), into machine readable formats and this includes the process of Optical Character Recognition (OCR) and design of digital fonts.

### **Current Trends and Innovations in Deep Learning Techniques for OCR in Low-Resource Script Recognition**

Optical Character Recognition (OCR) is a process that converts images of orthographic scripts to machine readable text. OCR is used for digitizing historical archives, helping language conservation.

OCR pipeline typically goes through multiple individual tasks including (i) Image acquisition - Extracting images containing text from multiple sources for offline images, and capturing live images for online extraction; (ii) Pre-processing - Application of image processing techniques, to increase raw image quality; (iii) Binarization - Isolation of text portions from the background visuals, for scenarios where text and images/videos are mixed; (iv) Layout Analysis - Dividing the images into regions (v) Segmentation - Partition of the image into pages, lines, words, characters and symbols; (vi) Feature Analysis - Identification and extraction of key features; (vii) Classification - Recognition of the symbols with scrip character-set; and (viii) Post processing - Use of pre-compiled vocabulary and language rules to auto correct the unrecognized words [39].

OCR engines generally use two approaches in the design of pipeline, namely (a) Supervised approach where labelled dataset is used to train the model and (b) Unsupervised approach where unlabelled data is used to build the model. Most unsupervised methods are semi-supervised using a small seed of labelled data for training the models.

### **Transfer Learning**

Transfer learning is a machine learning (ML) technique where a model trained on one task is repurposed or adapted for a second related task. Instead of starting the

learning process from scratch, transfer learning leverages knowledge gained from solving one problem and applies it to a different but related problem [94, 95].

In transfer learning, the pre-trained model (the model trained on the original task) serves as a starting point. Then, either the entire pre-trained model or some of its layers are retrained on the new task using a smaller dataset. This approach can be particularly useful when the new task has a smaller dataset, as it allows the model to generalize better and achieve improved performance by leveraging the knowledge learned from the original task.

Transfer learning is popular in natural language processing (NLP) tasks, as it enables faster model development and deployment, especially in scenarios where collecting labeled data for training a model from scratch is challenging or expensive.

### 4.3 Analyzing Deficiencies and Emerging Challenges in Low-Resource OCR Models

While there are plenty commercial and open-source OCR engines available for contemporary documents, however, low (or ultra-low) resource scripts differ in their requirements and hence off-the-shelf OCR engines cannot be put directly to use. The challenges for running OCR methods for low resource scripts can be illustrated as below.

- **Non availability of large volume of digital data:** Most OCR frameworks are designed with data hungry, deep learning classifiers under the hood. These classifiers require large datasets and hence cannot be used to Low Resource Scripts (LRS) effectively. Use of generative methods to build synthetic data also does not work since there is no initial labelled data for scripts which are not on digital footprint at all.
- **Constraints of human annotation due to limited user base:** Most of the ultra low resource scripts have a very limited user base, and these users are not on digital ecosystem. Availability of human annotators is scarce and methods like crowd-sourcing cannot be applied.

Application of transfer learning to build an OCR engine for low-resource script where a pretrained models for a high resource language is used as a foundation model and few shot learning is applied to customize it for low resource script suffers from the issues as illustrated below.

- **Training Bias:** Use of transfer learning using foundational model for some other scripts creates training bias for the foundational script. This happens



since the volume of data used for foundational model exceeds the one for low resource language by significant margin.

- **Catastrophic interference:** In case large volume of synthetic data is generated for low resource scripts and used on foundational models, there is a high chance of foundational model parameters getting interfered by training and defying the purpose of using foundational model.

## 4.4 VOLTAGE: Proposing a Novel and Versatile OCR Methodology for ultra low-resource scripts

We propose an automated versatile unsupervised OCR methodology (VOLTAGE) for very low resource scripts to address the gaps as illustrated earlier.

### 4.4.1 Strategic Selection of the Endangered Scripts for Experimental Focus

The primary purpose of our study is to architect an OCR methodology to streamline digital on-boarding for scripts with little-to-none digital resources. Therefore, the choice of the script(s) for experimenting, designing and validating the proposed methodology is a critical decision. We use *Takri* as our primary script (along with four additional scripts to generalise the methodology and make it agnostic of the underlying script), characterized by severely limited digital resources, absence of a labeled dataset, and a minimal user base.

George Abraham Grierson (1851–1941), linguist, and philologist, best known for his monumental work "Linguistic Survey of India." describes *Takri* and its variations as a script with shared inherent characteristics consequently classifying it as a "class of scripts" rather than a single script [96, 97]. *Takri*, like most Indic script falls under *Abugidas* class of writing systems [98], and some of the salient characteristics are summarised below:

- The character set of *Takri* comprises of 11 vowels, 33 consonants and 10 numbers.
- There are 10 vowel modifiers which can occur on the top, below, left or right of the consonants.
- *Takri* script does not contain headline unlike other Indic scripts like Devanagari
- Half forms are not used in most versions of *Takri*.

- Ligatures are also infrequently written.
- Most characters consists of connected components only.
- Compound characters are not present in Takri.

In order to make sure that the proposed methodology is generic and works across all other Indic scripts in similar capacity and accuracy, we also augment the experiment with three other scripts namely Modi (which was earlier used to write Marathi), Ol Chiki (which is a more modern but remote script to write Santali) along with Wancho (to write Wancho language used in north eastern regions of India). We also test the methodology with high resource language Gujarati (in both high and low data settings) to complete the experiment. While we discuss the results for all these scripts later, the purpose of using a variety of scripts is to establish the value of our work in more universal use.

## 4.4.2 VOLTAGE: Pipeline Design for ultra low resource scripts

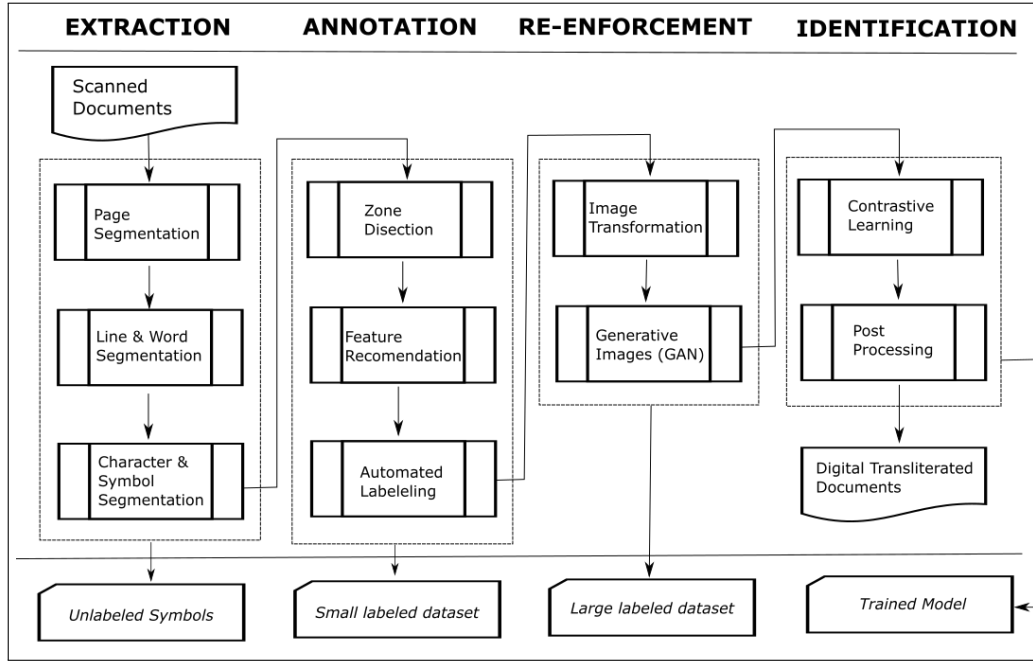
We propose VOLTAGE (A Versatile contrastive learning based OCR methodology for ultra Low resource scripts Through Auto Glyph feature Extraction), an OCR methodology using contrastive learning, GAN-augmentation and auto glyph feature extraction to enable digital on-boarding for near extinction scripts.

As illustrated in Figure 4.1, VOLTAGE follows the pipeline of tasks logically grouped into four parts, (i) Extraction - which comprises of pre-processing and segmentation; (ii) Annotation – includes automated feature engineering and unsupervised labelling; (iii) Re-Enforcement – including synthetic data generation using augmentation methods; and (iv) Identification - including classification and post processing tasks.

### Extraction

Prior to utilizing the input source for subsequent OCR tasks, it is imperative to undertake the extraction and segmentation of the text into distinct units, encompassing lines, words, characters, and symbols. Most OCR engines leverages computation of horizontal and vertical projections (HX and VY as illustrated in Eq. 4.1 and 4.2) to identify line separators and word separators (valleys points in HX & HY) [99, 100, 101].

$$VY_i = \sum_{x=0}^{x=Width} (\text{No. of black pixels for } x_i) \quad (4.1)$$



**Figure 4.1:** High level design for Versatile unsupervised OCR methodology for Low-resource scripts Through Auto Glyph feature Extraction (VOLTAGE)

$$HX_i = \sum_{y=0}^{y=Height} (\text{No. of black pixels for } y_i) \quad (4.2)$$

( $VY_i$  is the cumulative sum of all black pixels on  $x$  axis from left to right at  $i^{th}$  index;  $HX_i$  is the cumulative sum of all black pixels on  $y$  axis from top to bottom at  $i^{th}$  index)

Further, the segmentation of words into constituent characters presents a more peculiar challenge, one that is contingent upon the script category in question. While *alphabetic* scripts typically follow a more straightforward segmentation process, the task becomes notably more complex when dealing with *abugidas* scripts. This complexity arises from the inherent overlapping nature of individual characters within abugidas, compounded by the seamless integration of vowel modifiers into the character glyphs. We propose enhancement to equation 4.2 in order to accommodate the overlaps and inherent characteristics of indic scripts, we have observed that these changes helps in reduction of around 3% errors during the segmentation process for Takri script.

Segmentation of words into characters is slightly more complex (for *Takri* and similar scripts ) due to close vicinity of characters and overlaps. To solve this issue, we have enhanced Eq. 4.2 and compute the enhanced horizontal projection (EHX) which applies additional penalty in downward direction for character segmentation (Eq. 4.3) since the overlaps in Takri (and similar Indic scripts) occur in the upper

parts. We have observed that this technique, helps in overall reduction of segmentation errors by 3% (we tried with multiple values of penalty ranging from .1 to 1, and found a sweet spot at 0.3). Ol chiki and Wancho have the characteristics of alphabetic scripts, hence we use HX for character segmentation for these scripts instead of EHX. The choice of character segmentation method is dependent on the category of script being processed.

$$EHX_i = \sum_{y=0}^{y=Height} (\text{No. of black pixels for } y_i) + (\text{Penalty Wt.} * y_i) \quad (4.3)$$

It is observed that abugidas scripts have more complex structural characteristics for individual characters and the relationship between consonants and vowels is more intricate. While alphabetic scripts, such as "Latin" (also referred as Roman in some literature), assign individual symbols to consonant and vowel sounds, and treat them as separate entities, with vowels often represented as standalone symbols. On the other hand, abugidas scripts, such as "Devanagari" (and other Indic scripts), exhibit a more complex structure (within individual characters) where consonant-vowel pairs, known as syllabic units or graphemes, are combined into single characters. Hence the approach to further segment characters into individual symbols (consonants and vowel modifiers) need additional steps.

VOLTAGE users a three step process to achieve this -

- The initial stage involves *Skeletonization*, a process that diminishes the character's thickness to a single pixel thereby achieving consistency in thickness regardless of variations in the input.
- The next process is *Connected Component Labeling (CCL)*, also known as Connected Component Analysis (CCA) or Blob Extraction, helps to identify and label connected regions (components) in a binary image. Most symbols (apart from very few exceptions) are connected and hence this helps to isolate the parts of characters into individual symbols [102].
- The vowel modifiers in case of Indic scripts can happen on the top, below or either side of consonants, hence in the last step we apply *Zone classification* to isolate regions and separate characters into individual symbols.

The end product of "Extraction" phase within the VOLTAGE pipeline is a collection of unlabelled symbols into a single pool. Since we are working on ultra low scripts the size of the pool is small and ranges within 6-15K range.

**Table 4.1:** Unsupervised clustering accuracy for various zones for various k-means combinations.

	<b>Upper Zone</b>	<b>Middle Zone</b>	<b>Bottom Zone</b>
Distribution	23%	70%	7%
No. of Labels	5	50	4
Accuracy (without GFRS)			
50 iterations	91%	61%	93%
300 iterations	<b>96%</b>	69%	<b>97%</b>
With GFRS	-	<b>96%</b>	-

## Annotation

Annotation of extracted symbols is a very critical activity for effective OCR design. With a limited pool of unlabelled symbols, the next step is to label each character appropriately so that the dataset can be used to train the classifier in downstream parts. Labelling by the use of human annotators could be the most simplest and best option however this option is not scaleable and for ultra low resource scripts there are not enough people who can help with this.

For "Takri", running the "Extraction" steps resulted in the collection of 14,000 unlabelled symbols. We also know that Takri consists to 50 consonant and vowel symbols in the middle zone, 6 vowel modifiers in top zone and 4 in lower zone making a total of 60 categories of distinct symbols.

Partitioning of the unlabelled image dataset into visually similar categories with no ground truth can be accomplished by unsupervised clustering techniques. Pre-trained image models can be used to leverage parametric memory in such cases, and can help in good quality clusters. We use "ImageNet" as pretrained model and perform symbol partitioning within each zone separately <sup>2</sup> [103].

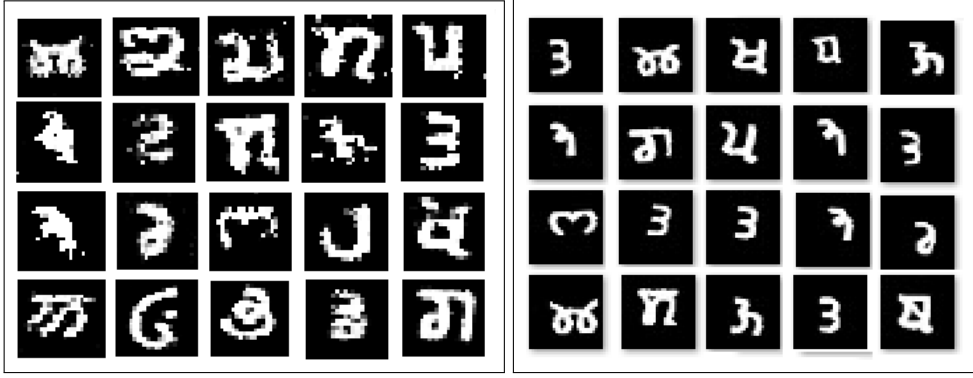
As illustrated in Table 4.1, while we run this for Takri, we observe that the accuracy of assigning the appropriate label in top and lower zone is 96% accurate, however the accuracy for middle zone characters was limited to 69% only. On further analysis we observe that as the number of clusters increase (which is proportional to the number of distinct symbols) the quality of partitions deteriorate. The linear dependency between number of classes and quality of partitions has also been observed in similar studies [104].

### Glyph Feature Recommendation System (GFRS)

We design a fully automated and universal procedure, which can run on any script to analyse the inherent glyph characteristics of script, recommend most appropriate

<sup>2</sup><https://paperswithcode.com/task/image-clustering>





**Figure 4.3:** Synthetic data generated for Takri - (Left Image) applying GAN algorithm and (Right Image) applying Image transformations

more sophisticated ML technique for generating new data samples that resemble a actual images. The key idea behind GANs is to have two neural networks, known as the generator and the discriminator, compete with each other in a game-theoretic framework. We illustrate some of the examples from this phase in Figure 4.3.

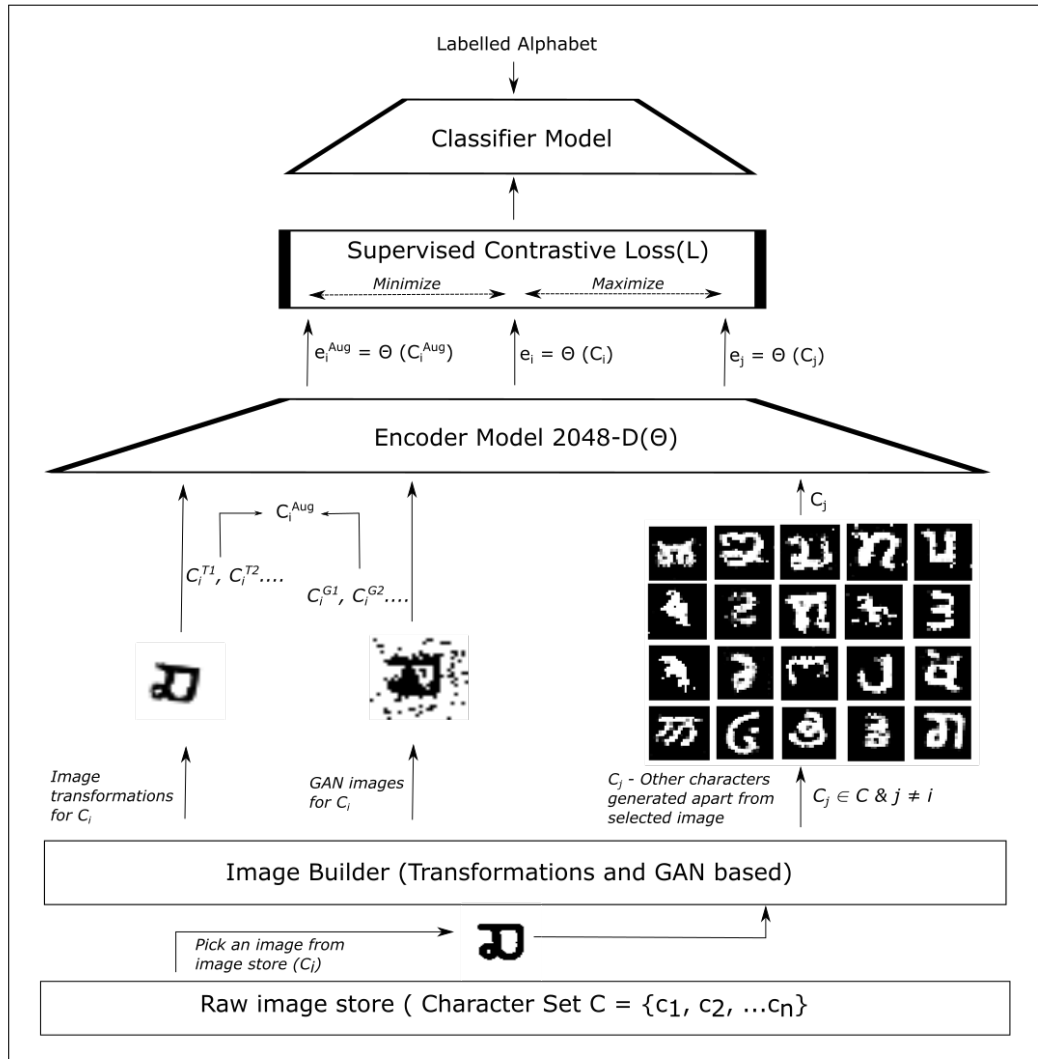
## Identification

Once there is large volume of labelled data available, the next step is to train a classifier which can accurately predict the character/symbol class. Traditional OCR engines typically use deep learning supervised methods including CNN [106] or a combination of CNN-LSTM [107] to train their models. VOLTAGE uses supervised contrastive learning as the classifier, leveraging augmented samples created in reinforcement phase. Trained GAN models help to generate positive and negative symbols which are fed into the encoder model. The encoder model encapsulates features and similarities of input pairs (both positive and negative) and maps them into a latent representation space. We apply supervised contrastive loss function from SupCon, to maximise the agreement between positive pairs (same character images) and minimize the agreement between negative pairs (different character images) [108].

$$L = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \quad (4.4)$$

Here  $z_i = Proj(Enc(\tilde{x}_i)) \in R^{D_p}$ , the  $\cdot$  symbol denotes inner dot product,  $\tau \in R^+$  is scalar temperature parameter. Index  $i$  is called anchor, index  $j(i)$  is called positive and other indices are called negative.  $P(i) \equiv p \in A(i) : \tilde{y}_p = \tilde{y}_i$  is the set of indices of all positive in multi-viewed batch (2N augmented samples, N being size of samples) distinct from  $i$ , and  $|P(i)|$  is its cardinality.

Figure 4.4 illustrates the end to end architecture for our classifier. We use Google colab for running our experiment. For GAN images we train for 150 epochs, 4 layer



**Figure 4.4:** Use of supervised contrastive learning for symbol classification

generator and network, adam optimiser and learning rate of 0.0002.

#### Post Processing:

Research has indicated instances where character misinterpretation occurs, from excessive character segmentation or inaccurate classifications. Consequently, the integration of post-processing techniques is crucial for rectifying these errors through language grammar and pattern analysis. VOLTAGE encompasses a repository of principles and guidelines for post-processing, facilitating error correction. While certain principles are universally applicable across scripts, others are tailored to specific script categories. These principles collectively enhance the overall accuracy of recognized text by incorporating linguistic context and syntax considerations.



**Table 4.2:** Empirical study for VOLTAGE on Takri on Machine Printed (MP) and Hand Written (HW) samples.

Zone	MP	HW
UZ (Upper Zone)	96%	88%
MZ (Middle Zone)	94%	85%
BZ (Bottom Zone)	97%	89%

## 4.5 Analytical Insights and Discourse on VOLTAGE’s OCR Results for Endangered Scripts

Empirical testing for an OCR pipeline can become ambiguous and subjective unless it is discussed, defined and correctly scoped [109]. We understand that VOLTAGE (like most other OCR engines) has a pre-processing step including segmentation of input image into pages, lines, words, characters and symbols before it can be further processed for recognition of characters.

OCR systems can be evaluated using multiple metrics including end-to-end (E2E) errors including all parts of OCR pipeline, or errors post segmentation phase (excluding the errors during segmentation phase, and those come under segmentation errors computed separately). Most work on OCR consider errors post segmentation phase which include Word Error Rate (WER) and Character Error Rate (CER), both of which are the ratio of mis-recognised entities to correctly segmented entities.

$$WER = \frac{N'_w}{N_w} \ \& \ CER = \frac{N'_s}{N_s}$$

where  $N'_w$  is count of mis-recognised words,  $N_w$  is correctly segmented words,  $N'_s$  is count of mis-recognised symbols, and  $N_s$  is correctly segmented symbols.

It is also noteworthy that each category of scripts (alphabetic, abjads, abugidas, logographic systems, and syllabaries) may have different definition of individual character/symbol. While alphabetic scripts (like English) use monolithic characters/symbols where each character is used in same form, however abugidas scripts (like most Indic scripts) use additional modifiers for vowels etc. which combine with base symbols and create additional glyphs [98].

For our work, we consider the root and marker symbols as separate while we compute CER. While we validate VOLTAGE for all the metrics on Takri, we use CER for validation across additional scripts.

We also conduct baseline studies to study the use of existing pre-trained models for high resource languages and its effectiveness on Takri using transfer learning. In

the end we conduct ablation studies for various components in VOLTAGE to establish the merit the proposed components within VOLTAGE pipeline.

## Applying VOLTAGE to Takri

The individual symbols in Takri (like any other abugidas category of scripts) can be split into three zones consisting of top and bottom zones (encompassing vowel modifiers), and the middle zone (primary symbol which may be a consonant or independent vowel). We calculate end to end errors (which includes segmentation errors along with classification errors) for each zone separately and illustrate the same in Table 4.2.

It is observed that the error metrics (E2E, WER and CER) are co-related, and hence most of the existing work conduct studies using the standard metric of CER. For further evaluation, we also use CER more extensively not including errors during segmentation and pre-processing [110].

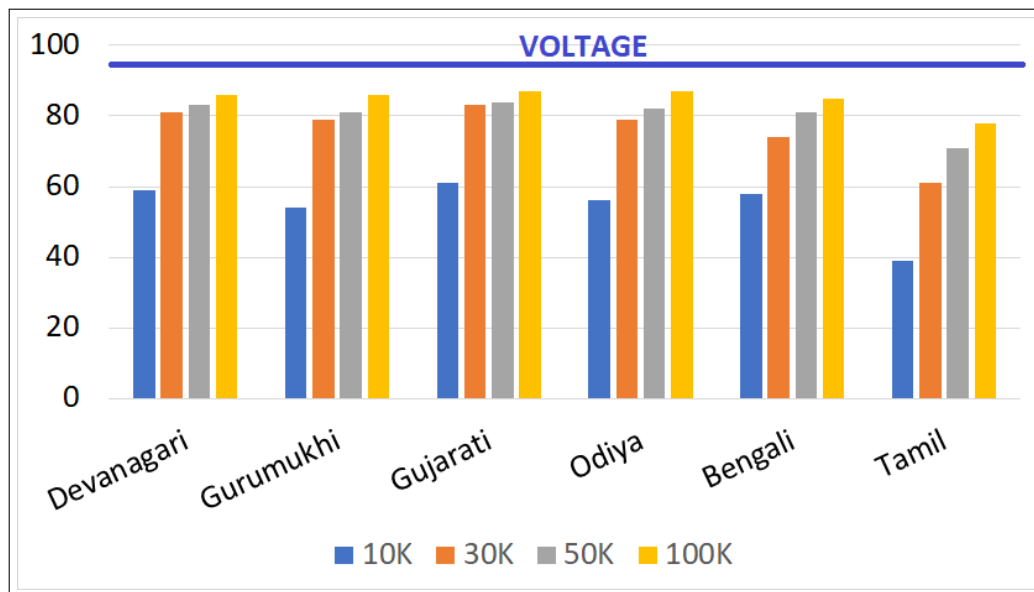
## VOLTAGE for other scripts

Table 4.3 illustrates performance of VOLTAGE pipeline across five scripts including Takri, Modi (historically used to write Marathi), Ol Chiki (to write Santali), Wancho and Gujarati. We use Gujarati in two combinations, initially in low resource setting taking a very small dataset and later as a high resource language. The purpose of doing this is to illustrate the effectiveness of our work in all forms and scenarios. We clearly observe that our work is generalised, agnostic of scripts being used and compares with SOTA accuracy wherever it is available. (For scripts where no benchmarks are available we have mentioned as NA in SOTA)

## Baseline studies

**Table 4.3:** Evaluation across other scripts. For Gujarati we experimented with two scenarios, (a) Gujarati LRL- Like low resource language and (b) Gujarati HRL- like high resource language

Script Name	Script type	SOTA Accuracy	VOLTAGE Accuracy	Dataset size	Glyph features
Takri	Abugida	NA	95%	14,051	9
Modi	Abugida	(84-94)%	93%	7,221	11
Ol Chiki	Alphabet	(83-92)%	91%	8,873	5
Wancho	Alphabet	NA	91%	6,500	8
Gujarati LRL	Abugida	(86-96)%	93%	7,643	9
Gujarati HRL	Abugida	(86-96)%	96%	200K+	9

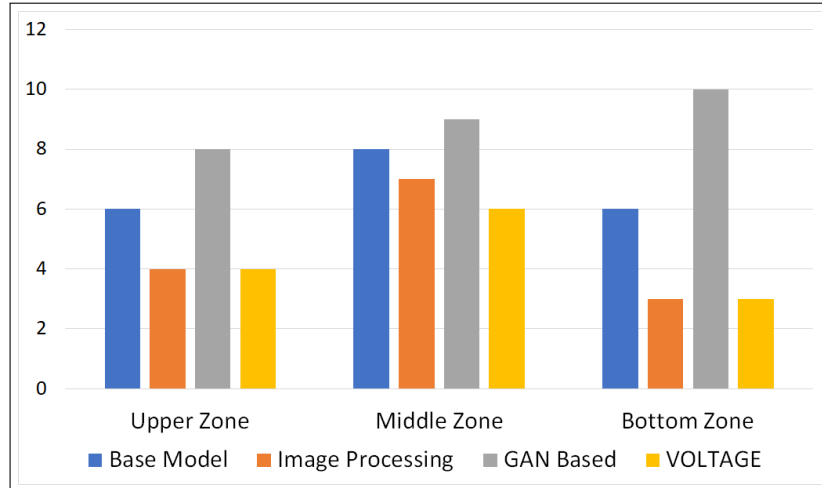


**Figure 4.5:** Use of foundational model and transfer learning using annotated Takri dataset

As illustrated in Figure 4.5, we use a variety of pretrained foundational models and apply transfer learning using the annotated Takri dataset created as part of our work. This baseline study identifies several notable issues with the use of pretrained model and use of transfer learning. Firstly, the selection of a foundational script is emphasized as pivotal, as demonstrated by the observed optimal performance when utilizing the Gujarati script in the case of Takri. However, it is noted that the process of selecting the appropriate foundational model entails a considerable degree of hit and trial (manual intervention), hence we refrain its integration into the overarching methodology. Secondly, the volume of transfer learning is deliberately constrained to a maximum of 100,000 samples. Scaling to higher sample sizes (beyond 100K) necessitates an extensive volume of data which is, (i) impractical for scripts with exceptionally low usage rates; (ii) risks of the phenomena of catastrophic interference, thereby undermining the fundamental objectives of employing a foundational model. An observation is also made regarding the convergence of most models to comparable outcomes as the sample size increases, however always falling short of those achieved by VOLTAGE.

## Ablation studies

In order to substantiate the effectiveness and contribution of key components in VOLTAGE we conduct ablation studies using Takri as an example. We already established and illustrated in Section 4.4.2 and Table 4.1, the use GFRS increasing the labelling accuracy by 27% (from the initial value of 69% to 96%). We further



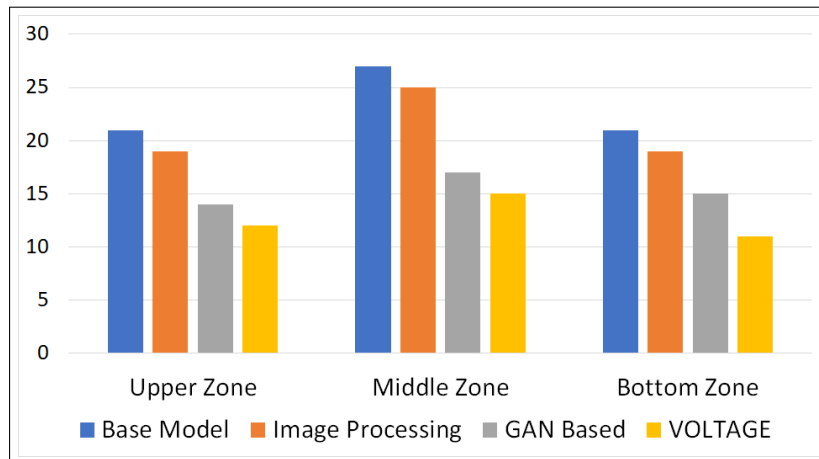
**Figure 4.6:** Character error rates (CER) for machine printed test scenario leveraging CNN-LSTM models and compare the results with VOLTAGE models

conduct ablation studies to substantiate the use of synthetic data (leveraging image transformation and GAN's) and the impact on accuracy.

Data is staged into three stages namely, (i) *Base dataset*: The initial small size raw dataset extracted from manuscripts etc. For Takri this is of the size of 14,000 labelled images; (ii) *Image Processing enhanced*: includes synthetic dataset created from Base images applying image manipulations including shear, brightness, restricted rotations etc. (iii) *GAN enhanced*: includes noisy synthetic images created by training GAN model for each individual character and using the same for creating noisy samples. We use data from each stage independently and use CNN-LSTM models to evaluate the effectiveness when data is consumed from each stage. We conduct these studies separately on upper zone, middle zone and bottom zone on machine printed and handwritten forms. As illustrated in Figure 4.6 and 4.7, the results clearly indicate that, while GAN models work better in handwritten scenarios due to noisy samples and image transformations work better in machine printed scenario, VOLTAGE works in both cases with improved accuracy beating both the methods in all scenarios.

## 4.6 Conclusion and Limitations

We present and empirically validate a comprehensive unsupervised OCR methodology, VOLTAGE, which incorporates an innovative Glyph Feature Recommendation System (GFRS) for efficient symbol labeling. VOLTAGE was developed using the Takri script, characterized by very limited resources, and its effectiveness and applicability were confirmed across various other Indian scripts. The proposed model



**Figure 4.7:** Character error rates (CER) for handwritten test scenario leveraging CNN-LSTM models and compare the results with VOLTAGE models

beats the state-of-the-art alternative models and achieves better accuracy for a variety of scenarios. This research has the potential to streamline the digitization process for ultra low-resource scripts, thereby facilitating their transition into digital formats and paving the way for the exploration of new use cases. Additionally, we have generated a sizable labeled corpus for the Takri script, which can be leveraged by fellow researchers with an interest in this particular script.

Although the current glyph feature repository is tailored for the stroke attributes of Indic scripts, its underlying concept can be applied to expand its utility to accommodate other script categories used in various other parts of the world.



# 5 | Dataset Generation Framework using Cross-lingual Embeddings and Content Based Image Retrieval

*“AI is the electricity of our era,  
and data is its fuel.”*

---

Andrew Ng

This chapter details the development of a novel data augmentation technique tailored for low-resource languages using newspaper articles (a ubiquitous source of linguistic data globally) [111]. The scarcity of annotated datasets poses a significant challenge for training robust deep learning models in these languages. To address this gap, we propose an automatic bilingual dataset creation technique leveraging newspaper articles as raw data.<sup>1</sup>

## 5.1 Introduction

The VOLTAGE model, as illustrated in Chapter 4, helps with the digital on-boarding of ultra low scripts to digital platforms, however most modern technologies powered by Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP) require sizeable data corpus for training state of the art models. With more than 7000 languages spoken worldwide, only a fraction of languages are supported by robust and sizable datasets. Consequently, a vast number of languages, particularly those categorized as low-resource languages, face substantial challenges in leveraging NLP techniques and tools due to the dearth or absence of high-quality parallel corpora [112].

Parallel corpora, comprising aligned text in two or more languages, play a crucial role in various NLP tasks such as machine translation, bilingual lexicon induction,

---

<sup>1</sup>Parts of this chapter have been published in the proceedings of SAC 2023 [111]

and cross-lingual information retrieval [113]. However, the scarcity of such corpora for low-resource languages severely hampers the development and application of NLP technologies in these linguistic communities. As a result, these languages are often marginalized in the AI technology landscape, limiting opportunities for communication, access to information, and participation in the digital economy for the native speakers.

We propose a novel methodology for extracting bilingual parallel corpora from newspaper articles, leveraging Content Based Image retrieval (CBIR) and multilingual embeddings. Our approach offers scalability and automation, addressing the labor-intensive nature of traditional corpus creation methods. The proposed methodology is agnostic of language combinations, and has been validated on a variety of low-resource languages.

We generate a sizeable Konkani-Marathi parallel corpus, comprising approximately 15,000 sentence pairs, and this dataset stands as the most extensive dataset available for this language combination, generated without manual intervention. This dataset holds significant promise for advancing research and development efforts aimed at enhancing digital enablement for digitally-niche semi-literates, by building AI models for low-resource languages.

We substantiate the value of the generated parallel dataset through a downstream task of machine translation, where we observe a notable improvement over the current baseline. This work, shall be a small step to contribute to the broader goal of fostering linguistic inclusivity in the digital era.

### 5.1.1 Defining Research Objectives: Multimodal AI for Automated Low-Resource Bilingual Parallel Dataset Creation

While the larger objective for this work is to build a parallel dataset useful for low resource languages, it is important to define and articulate specific objectives for the methodology to be designed.

- **General purpose, agnostic of the language of interest:** The algorithm should be versatile, general-purpose, robust and resilient, and adaptable to any language combinations.
- **Scalability, Efficiency and Quality:** The algorithm should offer improvements in efficiency and scalability compared to traditional methods of parallel corpus creation. The aligned dataset should maintain high levels of accuracy



and quality, ensuring that the aligned texts are linguistically accurate and suitable for use in various downstream NLP tasks.

- **Coverage, Diversity and Accessibility:** The parallel datasets should contain a wide range of domains and text genres. Since the purpose of our research is around semi-literate enablement, our objective should be to have sentence pairs that are representative of day to day usage of language, and simple to comprehend.

## 5.2 Survey of existing AI-Powered Multimodal Strategies and Identifying Gaps for Generating Bilingual Parallel Datasets in Low-Resource settings

Bilingual parallel data corpus creation, where one of the language is a low resource language has not been exhaustively worked on before. However it has been observed that the corpus mining techniques, depends on the factors such as the languages involved and the availability of resources (linguistic as well as pre trained computational models).

### Computational Methods

These methods leverage computational algorithms, including Artificial Intelligence (AI), Machine Learning models (ML), and Natural Language Processing (NLP) techniques to extract, process and align text collected from various sources, and align them to build the parallel text pairs.

#### **Web Crawling and Scraping for raw data extraction:**

This method helps to collect and extract text from multiple sources in multiple languages. Tools and libraries such as Scrapy and BeautifulSoup can be used to crawl websites and extract parallel text pairs from web pages.

#### **Word embeddings:**

Word embeddings represent individual words as dense, low-dimensional vectors in a continuous vector space. Word embeddings capture semantic relationships between words and are capable of capturing similarities, analogies, and word contexts. These embeddings become more enriched, when context is encapsulated. State-of-the-art models such as Embeddings from Language Models (ELMo), Generative Pre-trained Transformer (GPT), Bidirectional Encoder Representations from Transformers (BERT) etc. are examples of contextual embedding models. When language

specific embeddings are not available, Multilingual Embeddings can help in cross-lingual mapping and build parallel corpora.

**Machine Translation (MT)** For languages where some initial seed of parallel text corpora is available, and a pretrained MT model is available, it is also possible to build parallel corpora for a possible different domain of sentences than what the MT model was trained on.

#### **Data Augmentation (DA)**

DA refers to the methods that aim to expand the diversity and volume of data, and can be very useful to expand the volume of datasets. DA schemes are classified into two parts, extractive and generative. While *Extractive approach* for DA uses unsupervised data from multiple sources, applying web crawling, voice transcriptions, document scanning etc., *Generative approach* uses artificially synthesised data using various kinds of text synthesis approaches [40]. Extractive DA can be done by (a) Rule-based technique (b) Interpolation technique and (c) Model-based techniques [41]. Generative DA on the other hand is achieved by (a) Paraphrasing (b) Noising and (c) Sampling based methods [42].

## **Human-In-The-Loop Methods**

These methods rely on human involvement for collecting, annotating, mapping, and validating parallel text data. They may involve a variety of techniques, all of which are heavily depended on human labor. It is advisable to use these methods when computational methods are not feasible due to resource constraints, domain specialisation or linguistic complexities.

#### **Crowd-sourcing and Human Annotation**

Through the use of platforms such as Amazon Mechanical Turk and CrowdFlower, human annotators can help with the alignment of text passages in different languages, aggregate the results to create parallel corpora. While this method ensures high quality alignments, however it is often time consuming.

#### **Aligned Document Collection**

It is also observed that some of the pioneering literature gets translated into almost all the languages using worldwide, and can be used as an "Aligned Document Collection" to extract and map sections/ paragraphs and sentences. These documents may include religious texts like Bible or Ramayan, or famous fictional literature or speeches spoken by world leaders. While this gives an already aligned textual sections (chapters or smaller parts) however this method would also need some manual intervention and could have scalability issues.

## 5.3 Critical Assessment of Limitations and Challenges in Multimodal AI Approaches

While the methods mentioned for mining parallel corpora are effective, they also come with several challenges and issues, as outline below.

- **Optimizing scalability while mitigating noise:** For low-resource languages, both computational and human intensive methods of dataset generation suffer from scalability and quality due to limited resources, like source of raw datasets, linguistic expertise, and text embeddings models.
- **Language and domain agnostic:** Parallel bilingual datasets when created with the purpose of language translation for people with niche language requirements, should focus on general domains and easy vocabulary. Building datasets for specialized domains or enriched vocabulary is not useful.
- **Responsible data collection and curation:** Mining parallel data including for low-resource languages could raise ethical concerns such as copyright infringement and data ownership. Issues such as privacy, consent, and confidentiality should be of prime importance and in compliance with relevant regulations and guidelines.

## 5.4 Proposed AI-Powered methodology for Low resource Bilingual Parallel Corpus Extraction

### 5.4.1 The Design Intuition: Image-Pivoted Article Mapping and Advanced Sentence Embeddings

As illustrated in Figure 5.1, we observe a prevalent practice in the newspaper content, specially among the regional newspapers for marginalized languages and users. In order to optimize on the resource allocation, these newspapers often reuse the photographs taken by photojournalists who are responsible for capturing images of various events. While the reporters and editors are different and re-write the articles based on local context and user needs, however the photographs are reused. This is even more evident in case of "multilingual press" providing news and information to speakers of different languages.

Leveraging this observation, we implemented a methodology that utilizes images as pivotal elements for the mapping of articles across various language versions.



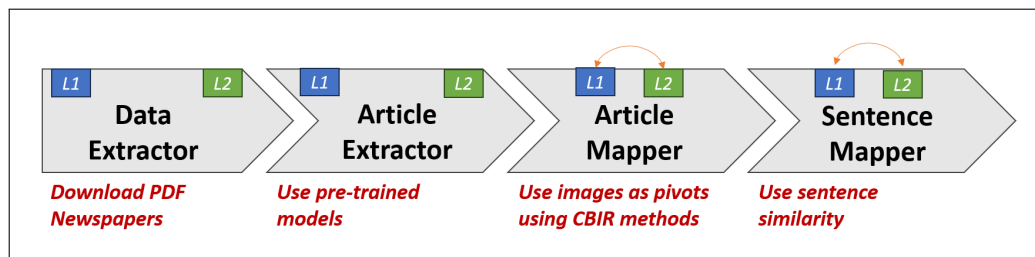
**Figure 5.1:** Article mapping using images as pivots.

Remarkably, our approach yielded a noteworthy achievement, achieving 100% accuracy in article mapping. We also observe that news content is highly diversified (as it contains information on multiple domains like politics, sports, entertainment, financial markets etc.) and also mostly free from copyright infringement and Personally Identifiable Information (PII) data.

Newspapers have a responsibility to handle PII data with care, especially when reporting incidents or events involving individuals. While reporting on news stories, journalists must balance the public's right to know with individuals' rights to privacy and data protection. It is mandatory for them to ensure compliance with relevant data protection laws and regulations, such as the General Data Protection Regulation (GDPR) in the European Union or similar laws in other jurisdictions. Hence sourcing data from newspaper is a responsible data collection method.

## 5.4.2 Manifesting the Design Philosophy: Experimental Setup and Implementation Details

As illustrated in Figure 5.2, our experiment runs in a sequential four phased approach, (i) *Data Extractor*, helps in collection of raw data from multiple newspaper sources and annotate with metadata (ii) *Article Extractor*, marks boundaries of individual articles and extract them into individual artifacts (iii) *Article Mapper*, uses the principles of content based image retrieval (CBIR) and use images as pivots to map articles in two languages and (iv) *Sentence Mapper*, as the last step where news content from mapped articles is mapped at sentence levels using cross-lingual language



**Figure 5.2:** Proposed data augmentation architecture. While Data Article Extractors (Data & Article) work on the two languages independently, the Mappers (Article & Sentence) use the languages in conjunction.

embeddings. We run our experiment on Konkani-Marathi combination in depth, and later validate for one more combination of low-high resource combination of languages.

## Data Extractor: Targeted retrieval of news content

Data Extractor facilitates the retrieval of news content from online sources<sup>2</sup>. However, for the choice of our language combinations, the downloaded files are in a non-machine-readable format, presenting the content as a static image devoid of searchable text layers. Additionally, a notable discrepancy exists in the size and length of newspapers across the two languages for the same date. To address these challenges, the downloaded newspapers are segmented into individual pages and meticulously labeled with pertinent metadata such as dates, page numbers, and language codes, making sure there is seamless referencing in subsequent downstream processes.

## Article Extractor: Annotate individual articles

Article Extractor, helps with two functions, (a) Marking boundaries for individual articles (b) Extraction of images and text (using OCR) within the marked article. Embedded articles are considered parts of parent articles in our work due to its strong association with parent article.

The information design in most newspapers is multi-column, to help with readability. Less words in a line supports less eyeball movements and make it easier to comprehend. Though this allows multiple articles, overlapping in a single view however marking boundaries for individual articles is a computationally complex task. We have used layout analysis dataset by Pattern Recognition and Image Analysis Research Lab (PRImA) for article boundary detection [114, 115].

An article is the smallest independent entity in a newspaper which contains complete information. Any further breakdown from article, would lead to incomplete

<sup>2</sup>For our experiment we use Bhaangarbhuin for Konkani and Goan Varta for Marathi as newspaper sources



**Figure 5.3:** Annotation nomenclature for various components in the article

packets of information entities and hence it is important to keep track of article components when broken down further.

We illustrate this further in Figure 5.3, and design annotation structure of multiple regions of interest (ROI), namely (a) Headlines ( $H$ ) (including sub headlines) (b) Images ( $I$ ) (c) Picture Captions ( $P$ ) and (d) Contents ( $C$ ). Indexations is performed within each category to make it structured. We further illustrate this mathematically, as below.

$$\text{Article}(a) \equiv \begin{cases} H_1^0, H_1^1, H_1^2 \dots & \text{Main article headlines} \\ H_2^0, H_2^1, H_2^2 \dots & \text{Embedded article headlines} \\ I_1, I_2, I_3 \dots & \text{Images} \\ P_1, P_2, P_3 \dots & \text{Picture captions} \\ C_1, C_2, C_3 \dots & \text{Content} \end{cases} \quad (5.1)$$

We leverage OpenCV for marking ROI and OCR for text extraction [116]. We have enhanced the OCR process by considering a combination of EasyOCR<sup>3</sup>, PaddleOCR<sup>4</sup> and Tesseract<sup>5</sup> and use majority voting for final decision. We have observed that using a combination of multiple techniques for OCR decreases the error by 3%.

For Marathi, we perform an additional step of spell-check and correction using

<sup>3</sup><https://github.com/JaidedAI/EasyOCR.git>

<sup>4</sup><https://github.com/PaddlePaddle/PaddleOCR>

<sup>5</sup><https://github.com/tesseract-ocr>



lookup method. Since there is no benchmark lookup for Konkani, we did not perform spell check for Konkani. We also observed that period(.) and comma(.) were misclassified in some scenarios (due to low quality raw data), creating issues with sentence boundary detection. We improved this by building a classification model on neighbouring words. Finally, each extracted article is stored as a individual text file, and each ROI within the article is labelled with markers as illustrated in Equation 5.1.

## Article Mapper: Map similar articles

Article Mapper, conducts comparisons between article images ( $I_i, I_j$ ) across the two languages ( $L1, L2$ ) to ascertain their similarity, considering the same date ( $dt$ ). Thereafter it constructs a collection of mapped article pairs ( $a_i^{L1}, a_i^{L2}$ ) when the image similarity score surpasses a predefined threshold. Embedded articles are mapped using headline content. We illustrate this further mathematically, to provide a formal and precise explanation.

$$\{(a_1^{L1}, a_1^{L2}), (a_2^{L1}, a_2^{L2}) \dots\} \equiv \theta(I_i^{L1}, I_j^{L2}) \quad (5.2)$$

$\theta$  is the image matching algorithm function

$I_i^{L1} \in \forall$  (images for date ( $dt$ ) & language ( $L1$ ))

$I_i^{L2} \in \forall$  (images for date ( $dt$ ) & language ( $L2$ ))

To the best of our knowledge, mapping of newspaper articles using pictures as pivots, has not been explored earlier in a similar context. It is however possible that similar images across language editions may vary in contrast, scale, illumination and orientation. It is therefore important to consider all these factor before scoring similarities between pictures.

As illustrated in Equation 5.2, given two images ( $I_i, I_j$ ), the goal of image similarity algorithm ( $\theta$ ) is to find out the degree of *similarity* of two images. One way of finding image similarity is by using traditional image processing techniques using local features in an image, commonly known as the *keypoints*. These keypoints are scale and rotation invariant, not affected by size and orientation of the image and can be used for image matching effectively (image processing algorithms like SIFT, SURF etc. uses keypoints for their processing [117, 118]). The other method is to use deep CNN based models. The CNN's automatically learn a representation of the image based on the objective function, provided data and the network architecture.

Though CNN based methods have an edge for more complex images where, however the images for our work are near similar and vary only in contrast and scale. We have observed that using SIFT along with OpenCV gives 100% article mapping accuracy, for our experiment [117]. We refrain from using more advanced forms of

image matching techniques using deep learning for our work for the ease of use and simplicity.

## Sentence Mapper: Map individual sentences

The extraction and alignment of data to the granularity of individual articles pose a relatively straightforward task. However, the sequential arrangement of sentences within the same article across the two languages is not guaranteed and hence the alignment of individual sentences becomes complex task.

We leverage a variety of semantic similarity methods (which we illustrate in detail in next section), and map sentences within the mapped articles ( $a_i^{L1}, a_i^{L2}$ ) to construct sentence pairs ( $s_j^{L1}, s_j^{L2}$ ) which can be considered as translations of each other. We illustrate this further in the below mathematical forms ( $\forall$  is used to represent "for-all" and  $\in$  represents "in").

$$\{(s_1^{L1}, s_1^{L2}), (s_2^{L1}, s_2^{L2}) \dots\} \equiv \delta(S_k^{L1}, S_k^{L2}) \quad (5.3)$$

$\delta$  is the semantic sentence similarity algorithm

$$S_k^{L1} \in \forall (\text{sentences for article } a_i^{L1})$$

$$S_k^{L2} \in \forall (\text{sentences for article } a_i^{L2})$$

Headlines and picture captions are mapped directly without any semantic similarity steps. We have observed that picture captions are generally are exact translations of each other.

## 5.5 Evaluation of proposed AI Model Performance on Multiple Language pairs

The empirical evaluation of the experiment is conducted into two aspects, (i) Intrinsic evaluation, where we ascertain the accuracy of mappings using sentence similarity metrics, and (ii) Extrinsic evaluation, where we train an AI model for machine translation task, and validate the accuracy of the model using standard metrics suitable for this task.

In addition to the use of Konkani-Marathi language combination, we also run our method on Punjabi-Hindi language pair to establish that general purpose of our method.

### 5.5.1 The choice of Metrics

#### Language Agnostic Sentence embeddings (LAS)



**Table 5.1: Semantic Textual Similarity (STS) for different sentence lengths. LaBSE Alignment Score (LAS), Sentence Length Alignment Score (SLAS) and Lexical Overlap Alignment Score (F-Score)**

Mapping Strategy	Short Sentences (1-10 words)	Medium Sentences (11-19 words)	Long Sentences ( 20+ words)
LAS	3.8	3.7	3.8
SLAS	3.4	3.4	3.2
F-Score	2.9	3.0	2.6

LAS converts sentence text into vectors to capture semantic information. We have used the pretrained LaBSE model which is based on BERT-like architecture and is trained on 119 languages of different origins (though Konkani is not one of these 119 languages) [119]. It claims to work universally including for the languages not part of its training corpus. It also claims to work for mining parallel data from the web for different language pairs with reasonable accuracy (BLEU scores of 35.7 for English-Chinese and 27.2 for English-German). We use LaBSE as one of the metrics, to compute the embeddings for Konkani and Marathi sentences independently and then compare the cosine similarity.

$$LAS = \frac{embedding_{Ko}.embedding_{Mr}}{\|embedding_{Ko}\| \|embedding_{Mr}\|} \quad (5.4)$$

### Sentence Length Alignment heuristics (SLAS)

We also explore SLAS to find the semantic similarity based on sentence length, and its position within the article. Sentence similarity is calculated between each sentence pair, and maximum similarity combination is extracted.

$$SLAS(s_i^{Ko}, s_j^{Mr}) = \alpha \left( \frac{(\lambda_i \phi(s_i^{Ko}) - \phi(s_j^{Mr}))^2}{\chi} \right) \quad (5.5)$$

In this equation  $\phi$  refers to the sentence length function in terms of number of words;  $\lambda_i$  is the sentence length ratio for articles  $a_i^{Mr}$  (which refers to the  $i^{th}$  article in Marathi) to  $a_i^{Ko}$  (which refers to  $j^{th}$  article in Konkani) to account for average length across two languages;  $\rho$  is the penalty factor to account for relative position of sentence in the article;  $\alpha$  is the normalization factor chosen so that the sum of SLAS metric for a fixed  $s_i^{Ko}$  over and all possible values of  $s_j^{Mr}$  is equal to unity [120].

We have used sentence lengths( $\phi$ ) as a function of number of words in a sentence, after filtering punctuation marks. We have observed that the average length of headlines and picture captions in our dataset is similar for Konkani and Marathi. This establishes that the length ratio for Konkani and Marathi is almost the same, hence

we don't have to apply any adjustment factor for sentence length across languages. However, the number of sentences within the article pair may not be similar, hence the ratio of sentence lengths within article ( $\lambda_i$ ) is used to find the relative position within the article.

**Lexical Overlap** In order to apply Lexical Overlap metric, we use English as a pivot language and perform lexical translation to English for both the languages. Marathi to English lexical conversion is a fairly simple task, using google translate. For Konkani, we use English-Konkani dictionary by Maffei and Xavier [121] for lexical conversion. The dictionary uses roman script for Konkani, hence an additional step of transliteration is also applied.

We then calculate precision, recall and F-Score score to measure the overlap statistically. Our results indicate that this method of evaluation is not accurate for our language combination, primarily because of the non standard Konkani-English digital dictionary along with change of script. (Refer to Table 5.1 and 5.2 for comparison)

$$\begin{aligned} Precision(P) &= \frac{count(overlapping\_words)}{count(Konkani\_words)} \\ Recall(R) &= \frac{count(overlapping\_words)}{count(Marathi\_words)} \\ FScore(F_1) &= 2 \cdot \left( \frac{P \cdot R}{P + R} \right) \end{aligned} \quad (5.6)$$

We use all the sentence mapping measures, for Konkani-Marathi combination and perform detailed empirical evaluation.

## 5.5.2 Empirical Evaluation

### Konkani-Marathi corpus evaluation

We analyse the quality of parallel corpus by applying human annotation on a subset of sentence pairs from our experiment. For defining the annotation scores, we use Semantic Textual Similarity (STS), characterised by six ordinal levels ranging from complete semantic equivalence (5) to complete semantic dissimilarity (0) [122]. STS is the most widely used evaluation metric for parallel data augmentation tasks, and used by multiple people working on this field [123]. Unlike surface-level similarity, which might be based on the number of shared words, semantic similarity focuses on the meaning conveyed by the text. Two sentences can be semantically similar even if they use different words.

**Table 5.2: STS for different article sizes (in terms of number of sentences within the article)**

Mapping Strategy	(1-5) sentences	(6-15) sentences	(16+) sentences
LAS	3.8	3.8	3.7
SLAS	3.1	3.5	3.3
F-Score	2.8	2.9	2.9

$$\text{STS Score} = \begin{cases} 5 & \text{completely equivalent, as they mean exacty the same thing} \\ 4 & \text{mostly equivalent, but with some trivial differences} \\ 3 & \text{roughly equivalent, but with minor differences in details} \\ 2 & \text{barely equivalent, conveying some related information} \\ 1 & \text{not equivalent, but on the same topic} \\ 0 & \text{completely dissimilar} \end{cases} \quad (5.7)$$

The annotators employed for our work were fluent in Konkani and Marathi and were compensated for this task. We have sampled a total of 900 sentences, in two phases. In the first phase we have sampled 200 pairs from each sentence alignment strategy (making a total of 600 sentence pairs) and analysed the results. In the second phase we have sampled, another 300 pairs for the most appropriate sentence mapping strategy as evident from first phase analysis. The sampling was stratified, shuffled randomly such that no ordering is preserved. We illustrate the experiment in two parts as below.

#### Phase 1:

We understand that multilingual embeddings capture the meaning of sentences across different languages in a shared vector space, are context-aware, go beyond surface level lexical forms and are robust across different domains. We observe the same findings in our experiment and as illustrated in Table 5.1 LAS consistently over scores other metrics and hence is most appropriate for this task. We further validate the metrics on articles of different sizes (since longer articles would be more complex to map) and find LAS as most appropriate again. Table 5.2 illustrates this result quantitatively establishing the fact that LAS is consistent and does not get impacted by size of articles.

We also compare the spearman correlation (also referred as  $\rho$ ) [124] between LAS and SLAS alignment methods for various article lengths. We discarded F-Score since the STS metric in this case was below 3 and hence not appropriate for

**Table 5.3:** Spearman correlation between LAS and SLAS for multiple sizes of articles

Article Size	Spearman correlation ( $\rho$ )
< 10 sentences	0.443
11-20 sentences	0.392
> 20 sentences	0.316

our work. Spearman correlation is a non-parametric measurement of the strength of monotonic relationship. This means that if one variable increases (or decreases), the other variable also increases (or decreases). Non-parametric correlation is less sensitive to outliers than is its parametric analog. We have observed that LAS and SLAS have moderately strong correlation for articles with less number of sentences. This also explains why LAS and SLAS have similar STS results for short articles (and short sentences). Refer to Table 5.3 for details.

$$\rho = 1 - \frac{6\sum d_i^2}{n.(n^2 - 1)} \quad (5.8)$$

where  $d_i$  is the difference between the ranks of the  $i^{th}$  pair of values and  $n$  is the number of pairs.

#### Phase 2:

Based on the results from first phase, we have used LAS as our final alignment method. Our Konkani-Marathi parallel corpus contains 14,448 sentence-pairs (aggregating a total of 28,896 sentences extracted). The average length of sentences for Konkani and Marathi are within the similar range due to morphological similarity of these languages. Table 5.4 illustrates the count and average sentence length across multiple headers for our corpus.

For second phase, we have evaluated the STS on a corpus of 500 sentences (200 from first phase LAS evaluation and another 300 in second phase). We have observed that the average STS in our corpus is 3.7 and more than 92% of our mapped sentences, have STS score of greater than 3 (Refer Table 5.5).

Our results indicate that our methodology is complete and reasonably good to

**Table 5.4:** Corpus sentence count and sentence length at multiple headers

Mapped Entities	Count	Average Length	
		Konkani	Marathi
Articles	12,107	14.99	15.72
Headlines	1,726	7.97	6.95
Picture Captions	615	14.53	13.96

**Table 5.5:** Cumulative Semantic Textual Similarity (STS) in final corpus

STS	Sentences Mapped
$\geq 4$	76.15%
$\geq 3$	92.18%
$\geq 2$	96.19%
$\geq 1$	100.00%

**Table 5.6:** Evaluation for translation task on Punjabi-Hindi combination

Metric Name	Value
Mapped articles	150
Mapped sentences	2200
STS Score	3.73

build a scalable model for parallel corpus generation. To the best of our knowledge our final dataset is by far, the largest parallel corpus for Konkani-Marathi and can be used for multiple applications.

## Punjabi-Hindi corpus evaluation

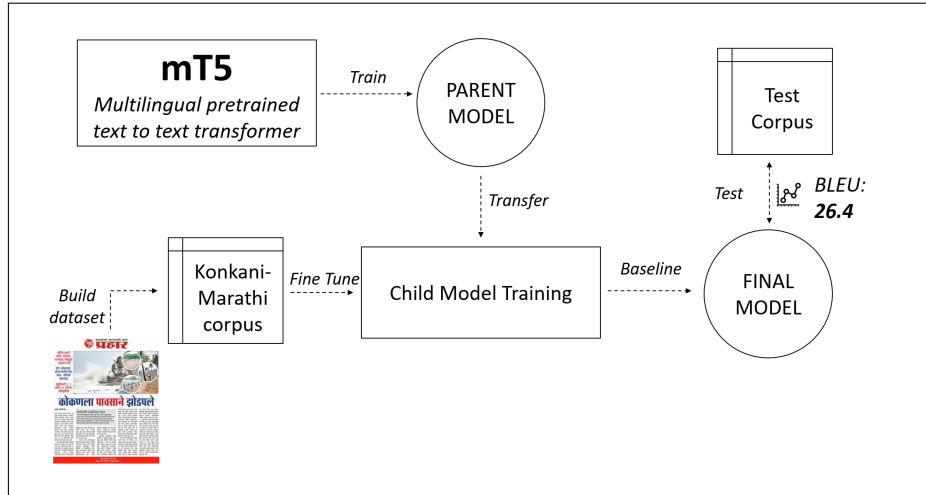
We executed the same experiment on Punjabi-Hindi language combination, following the exact same steps. We used Ajit<sup>6</sup> as our source of raw news corpus for Punjabi and Hindi news. We used a smaller set of news articles across one month's time span. Our intent of doing this as an extended experiment was to analyse and validate the significance of our work on other language pairs. Table 5.6 illustrates that the results on this language pair is at par with our primary experiment which clearly indicates that our methodology is universal and can be used to build large size corpus for a variety of languages.

### 5.5.3 Case Study

We already did a manual evaluation to find out the quality of mapped sentences using our methodology. However the evaluation is based on analysis of a small subset of sentences and may not be enough to validate the quality of entire dataset. Therefore to access the quality of the entire Konkani-Marathi aligned corpus, we perform extrinsic evaluation using the downstream machine translation task which shall leverage the entire aligned corpus as an input.

It has been established that cross-lingual transfer learning can help in improving accuracy of translation tasks for LRLs when the size of available corpus is limited

<sup>6</sup><https://www.ajitjalandhar.com/>



**Figure 5.4:** Extrinsic evaluation of dataset using Machine translation task

[125]. This method works by using a resource rich language combination (the parent model), then transfer some of the learned parameters to the low-resource pair (the child model) to fine tune the parameters. It has also been established that using pre-trained multilingual sequence-to-sequence models as parent models, helps for data scarce language combinations [126].

As illustrated in Figure 5.4, we use mT5 (multilingual pre-trained text to text transformer) as our parent model, and further fine tune with our Konkani-Marathi parallel corpus (headlines and article content) [127]. We have used picture captions as ground truth to test our translation model and achieved the BLEU score of 26.4. In an earlier study where a small manually annotated data corpus by Indian Languages Corpora Initiative (ILCI) was used to build a Neural Machine Translation (NMT) model (sequence to sequence encoder-decoder transformer architecture) for the same language combination, a BLEU score of 23.5 has been achieved [128]. Our work has an improvement by around 3 BLEU points.

## 5.6 Conclusion and Limitations

In this chapter we present a simplistic novel methodology for parallel data generation using universally available newspaper articles. We empirically validate our method for multiple language combinations and find it highly effective. Our Konkani-Marathi corpus is largest available corpus which has been created without human intervention. We achieve an improvement of 3 BLEU points on existing benchmark.

While the method described should in-principle work for all language combinations where newspapers are in circulation, however it can be constrained for languages where sufficient OCR tools/packages are not available. And our last chapter

addresses that limitation to a large extent, by the use of VOLTAGE framework.





## **Part 4**

# **Enabling "Digitally-Neglected" Semi Literates**



# 6 | Generative AI powered Multimodal Semantographic Metalanguage

*“I dream of a Digital India where access to Information knows no barriers.”*

---

Narendra Modi

This chapter illustrates the design of a visual metalanguage for effective communication in semi-literate settings, where exposure to orthographic scripts is bare-minimum [129]. While the orthographic forms of writing systems focuses on semantics and syntax, using a vocabulary of concepts, the visual forms of communication uses imagery as a communication medium. Our model limits the size of ideographic vocabulary making it easy to use, is extensible to new domains and can work universally.<sup>1</sup>

## 6.1 Introduction

In today’s increasingly digital world, the use of technology has become ubiquitous, permeating almost every aspect of daily life. However, amidst the rapid advancement of digital devices and services, there exists a significant segment of the population that faces unique challenges in navigating this digital landscape: individuals with bare-minimum academic literacy skills. These individuals possess elementary knowledge to decipher numbers, and read simple words, however lack proficiency in reading large portions of text. For them, the use of smart phones presents formidable hurdles that impede their engagement in the digital world. We refer these groups of individuals as "Digitally-Neglected" in our work [10].

---

<sup>1</sup>Parts of this chapter have been published in the proceedings of SAC 2023 [129]

In recognizing the importance of inclusivity and equitable access to technology, it becomes imperative to develop solutions that empower and enable digitally-neglected semi-literates to harness the benefits of digital connectivity. Architecting AI-models for bridging the digital divide, not only promotes the social inclusion but also unlock the untapped potential of a substantial portion of the population, fostering greater opportunities for learning, communication, and economic empowerment. Therefore, creating technology tailored to the needs of these individuals is not just a matter of convenience; it is a fundamental step towards building a more inclusive and accessible digital society [57].

In this chapter, we introduce a unique and universally applicable ideographic metalanguage known as "Multimodal And Global Ideographic Code with Hierarchical Ontology for Digital platforms" (MagicHood). MagicHood represents an innovative communication approach that surpasses academic, linguistic and cultural boundaries. It utilizes ideographs, also known as pictographs, combined with textual elements. The proposed multimodal approach combines the power of visual communication along with morphology of orthographic scripts, enabling digital literacy for semi-literate individuals who have been historically neglected in digital environments.

We employ both quantitative and qualitative testing methodologies to validate our approach. Quantitatively, we assess the effectiveness on semantic comprehension and gauge user satisfaction and the utility of our work using various metrics. To the best of our knowledge, there is no other universal multimodal-ideographic metalanguage for semantic simplification which has been applied for semi-literate enablement. The research presented in this chapter represents a potentially significant milestone, with the potential to be expanded for a wide range of applications beyond those outlined herein.

### 6.1.1 Articulating Objectives: Multimodal Generative AI Metalanguage for Users with Limited Orthographic Script Familiarity

While the larger objective for this work is enable digitally-neglected semi-literates using AI-models, it is important to define and articulate specific objectives for the proposed methodology and keep that as north star for the work.

- **Universal Adoption:** The proposed AI-model should have an universal application and should be agnostic of languages, academic qualifications of end users, domain versatility and applicability across various geographies, social and cultural backgrounds, making it a global metalanguage.

- **Reduced learning curve:** The proposed method should be easy to adopt for faster on-boarding, reduced errors and faster time to value. Since the intended recipients of proposed work are digitally-neglected semi-literates who suffer from digital anxiety it is very important to build a system with minimum learning efforts.
- **High Engagement:** The proposed model should be interesting to use, and have a good user experience, leading to higher user adoption and overall satisfaction. It should also harness the potential of digital ecosystem and create an engageable human experience.

## 6.2 Survey of current Ideographic Methods of Communication on Digital Platforms

Most existing research on multimodal communication leveraging ideographic communication can be categorised into two categories (i)"Static Systems" where formulation of ideograph is accomplished by rule-based extraction from a predefined knowledge base based on context, and (ii) "Dynamic Systems" where the synthesis of the ideograph is done either by careful assembly of set of images from knowledge base or open internet or generative methods. These systems are open ended and hence have no restrictions on what images can be suggested. Appendix D illustrates some examples from these methods.

### Static and Extractive Visual Communication Methods

Most linguistics refer *Static Methods* as **Semantographics** where the expression of semantics is accomplished with predefined ideographs than using orthographic forms. It is being observed that when these ideographs are properly designed, they seem more natural, and easier to learn and use, require less memorising and results in fewer mistakes[130].

While there have been multiple attempts made to build ideographic methods of communication, some of the pioneering works can be mentioned as below.

The oldest form of ideographic communication which is still in use is the use of **Naxi** script. NaXi ideographs belongs to the NaXi language of Yi language branch of Tibetan-Burmese languages which was created by the ancestor of NaXi [45]. **Iso-type**, also being referred to as the Vienna Method of Pictorial Statistics was developed in early 1920's uses a graphical form within a two-dimensional syntax to show social, technological, biological and historical connections [46]. **Blissymbolics**, was another attempt in the same era consisting of a set of symbols that represent basic

objects in the world and their features. **Nobel Language** is recent attempt to build an ideographic language consisting of 120 basic signs and many arrows of different shapes designed for cross lingual communication [48].

In the recent times, with the advent of digital devices some attempts were made to fine tune the use of ideographs for digital communication. **iConji** is one such example which is designed for communication on digital platforms, as an extension of emojis with built-in translations to twelve languages [49]. There is also in-progress work happening referred to as **Able to Include** project which seeks to improve the living conditions of people with intellectual or developmental disabilities (IDD) using Augmentative and Alternative Communication (AAC) [50].

## Dynamic Visual Communication Methods

There are a variety of Text-To-Picture (TTP) methods developed using computational methods, however we can broadly classify them into Pre Generative-AI era and Post Generative-AI era.

**Pre Generative-AI:** The earlier forms of generative methods follow similar processing pipeline, (i) Text analysis and dependency parsing, along with selection of words which are pictureable (can be imagined visually) (ii) Querying knowledge base for visual representations of pictureable words and assembly of extracted images to the final output (as a single image or set of images). Some of the recent works on this include **Illustrate It**, **AraTraductor** and **Word2Image** [131, 132, 133, 134].

**Post Generative-AI:** In the recent years, the rapid pace on the development of Large Language Models (LLM) (along with multimodal LLM) and public release of tools such as ChatGPT/LLAMA has attracted wide attention, optimism and concerns [52, 53, 54, 135, 136]. These models are capable of generating new content, in forms of text, audio, visuals etc. understanding the patterns in existing data, and hence is referred to as "Generative-AI". There are multiple text to picture pretrained transformers available today including, **Dall-E** by OpenAI/ Microsoft, **Imagen** by Deepmind/ Google, **DreamStudio** by Stability AI etc. [137, 138, 139].

### 6.3 Uncovering Challenges and Opportunities in Multimodal Communication Methods

While the methods mentioned work for some use cases, applying them to build a formal metalanguage for semi-literate communication, which can be used by people with negligible academic education has not be explored. We illustrate some of the gaps in existing methods, which are addressed in the proposed work [93].

- **Reduced user attention:** The inventory of symbols generally becomes very large, resulting in high acquisition time and reduced user engagement. For dynamic methods (specially using GenAI) the end result is non-deterministic and hence it is difficult to build knowledge and user engagement.
- **Extensibility for new ideas and concepts:** With a flat/limited structure of symbols and operators, the creative use of creating new concepts using existing symbols is restricted.
- **Lack of syntax:** Human interpretation of set of images in isolation is unintentionally influenced by the native language syntax of its users, hence a bare-minimum native text (along with ideographs) can greatly help in providing grammar and reduction of misinterpretation.

## 6.4 MagicHood: Introducing Generative AI Powered Methodology for Hierarchical Ideographic Communication on Digital Platforms

We establish our design philosophy on three principals, which we infer from the survey observations as illustrated in Chapter 2. The proposed model should follow (i) Minimalism, for reduced complexity, on-boarding duration and ease of use, (ii) Being versatile and expandable, to enable extensions to other domains and concepts making it universal for use, (iii) Digital-First, leveraging the power of digital technologies specially for the design of user experience.

We leverage the principles from the linguistic theory of Natural Semantic Metalanguage (NSM) establishing universal semantics of all languages, decoupling all concepts into elementary semantic concepts (referred to as "Semantic Universals" in NSM) [19, 20]. We have discussed NSM and its application for semantic simplification in Chapter 3.

### 6.4.1 Strategic Integration of Ideography and Text in Multimodal Communication

We analyse the text message dataset sourced from over 300 semi-literate respondents, collectively containing approximately 3000 messages in English [55]. Distribution on Parts of Speech (POS) within this dataset reveals that nearly two-thirds of the lexicons are comprised of nouns and verbs, thus establishing them as the most prevalent POS categories in the digital communication text. We further validate these findings

by analysing the Sketch Engine, specifically the English Web 2020 corpus, which encompasses a vast collection of 36 billion words and observe that 60% of the words within it are nouns and verbs [140, 141]. Accordingly it is established that two-thirds of the text communication comprises of nouns and verbs only.

Furthermore, we analyze the average complexity of words across various parts of speech, using a set of readability scores. We observe that nouns and verbs emerge as the most intricate components, difficult to read/comprehend within a sentence (adverbs, adjectives, pronouns, and conjunctions exhibit a comparatively simpler linguistic complexity).

It is thereby concluded that nouns and verbs are most abundant and complex semantic elements within the digital communication text, which are difficult to read and understand. And hence we spoke the ideographic simplification for these elements in our work.

## 6.4.2 Mathematical Model and Ontology

### Mathematical Model

The proposed AI-models leverages the linguistic framework of Natural Semantic Metalanguage (NSM) to simplify complex entities into simpler semantic universals, applying the hierarchical semantic structure of fundamental elementary concepts. The proposed methodology adheres to a three-tier hierarchy, which is inline with NSM linguistic theory.

- On the highest level, the semantic concepts are grouped into categories that share a common semantic core or meaning element. These constructs are referred as *Semantic Classes (SC)*. Examples include "Human" (comprising human semantic elements like parts of body, relationships etc.) and "Event" (comprising of social events, official work events etc.).
- The second level denotes more specific patterns or nuanced distinctions within Semantic Classes and can be described using a similar set of explanations (referred as explications in NSM theory). These constructs are referred as *Semantic Templates (ST)*. For illustration purpose, "Human Relationship" (containing concepts like father, mother, brother etc.) can be a semantic template within the Semantic Class of "Human".
- The third level comprises elementary concepts which are used as key-value pairs to explain/describe individual elements within semantic template. These



pairs of constructs are referred to as Semantic Variables (SV) and Semantic Molecules (SM). Using the same example the concept of "Father" can be described as  $\{(Gender=Male)(Path=Parent)\}$ . Here the constructs of Gender and Path are Semantic Variables and Male and Parent is Semantic Molecule.

Mathematically, in a closed form the hierarchical breakdown can be illustrated using the equations as below.

Assume:

$$\begin{aligned} \text{Semantic Class}(SC) &= \{sc_1, sc_2, \dots, sc_{n1}\} \\ \text{Semantic Template}(ST) &= \{st_1, st_2, \dots, sc_{n2}\} \\ \text{Relationship}(R^{sc-st}) &= x, \{Y\} \mid x \in SC \ \& \ Y \subset ST \end{aligned} \quad (6.1)$$

Semantic ideas within each ST can be depicted using multiple combinations of (sv,sm) pairs. The conversion of ST into a collection of Semantic Variables (SV) (as keys) and Semantic Molecules (SM) (as values) tuples is initially carried out manually when we build the initial ontology. Later the ontology is used as a context, to generate new combinations using Large Language Model (LLM).

$$\begin{aligned} \text{Semantic Variable}(SV) &= \{sv_1, sv_2, \dots, sv_{n3}\} \\ \text{Semantic Molecule}(SM) &= \{sm_1, sm_2, \dots, sc_{n4}\} \\ \text{Entailments}(E) &= \{e_1, e_2, \dots, e_{n5}\} \\ \text{where } e_i &= \{(sv_i^1, sm_j^1), (sv_i^2, sm_j^2) \dots\} \\ &\forall sv_i \in SV \ \& \ \forall sm_j \in SM \\ \text{Relationship}(R^{st-e}) &= x, \{y\} \mid x \in ST \ \& \ y \in E \end{aligned} \quad (6.2)$$

Using the mathematical model as described, we construct a two step process to identify and simplify complex concepts into simpler elements, (i) Segment the sentence into its constituent parts, perform parts of speech (POS) classification, and identify ideographical concepts; (ii) Describe the concepts, into hierarchical semantic flow, into SC,ST, SV and SM pairs.

STEP 1

$$\begin{aligned} \text{Sentence}(S) &= \{w_1, w_2, \dots, w_{n'}\} \\ \text{Complex Words}(W) &= \{w_{n1}, w_{n2}, \dots, w_{nn''}\} \\ W &\subseteq S; \\ \forall POS(w_{ni}) &\in \{Noun, Verb\} \end{aligned}$$

**Table 6.1:** Examples from our dataset, following our mathematical model

Concept	Semantic Class	Semantic Template	Semantic Variable & Molecule Pairs
Mother	Human	Human Relationship	(Path,P) (Gender,F)
Nephew	Human	Human Relationship	(Path,S, C) (Gender,M)
Grandfather	Human	Human Relationship	Path,P; P) (Gender,M)
Wedding	Event	Private Event	(Category,Matrimony)
Tomorrow	Time	Temporal	(Marker,Day+1)

STEP 2

$$\forall(w_{ni}) \in (sc_j, st_k)$$

$$\forall(st_k) \in \{(sv_1, sm_1), (sv_2, sm_2), \dots, (sv_l, sm_l)\} \quad (6.3)$$

Table 6.1 illustrate some example to further clarify the process of semantic simplification.

## Building Initial Ontology

It has been observed that semantic association between two concepts can be most appropriately evaluated via dense vector representations [142]. These vectors can be leveraged as a reference to identify and baseline the hierarchical semantic categories (SC, ST). There are generally two categories of these vectors, (i) Context independent vectors like FastText, Word2Vec etc. and (ii) Transformer based context aware vectors like ELMO and BERT [143, 144, 145]. We use our dataset (described in Chapter 2) and form semantic clusters using a variety of vectorization methods along with clustering algorithms (hierarchical, agglomerative, graph based etc.). As illustrated in Table 6.2, we observe the best performance using a combination of BERT (for embeddings) and BIRCH (for clustering) [146, 147, 148].

Clustering helps us to form groups of semantic classes (SC) and semantic templates (ST) as top two level of semantic structures. Optimization on the count of

**Table 6.2:** Semantic clustering approaches and findings. Human evaluation is normalised on 0-1 scale with 1 being best score. (W2V - Word2Vec, FT-FastText)

	W2V	FT	ELMO	BERT
K-Means	0.78	0.84	0.89	0.92
DBSCAN	0.79	0.78	0.90	0.91
<b>BIRCH</b>	0.81	0.79	0.93	<b>0.94</b>
Agglomerative	0.88	0.89	0.72	0.93

SC/ST is accomplished by minimising intra cluster errors and maximising inter cluster errors using Sum of Square Error (SSE).

The next step on the further breakdown of semantic templates into semantic variables and molecules is conducted by human annotators, to create the initial ontology. This ontology is later used as a context to automatically generate more concepts using Large Language Model (LLM) as an inferencing engine. The baselined initial ontology for our dataset with exact count, and hierarchy is illustrated in Figure 6.1.

### 6.4.3 Implementing Proposed Mathematical model for Multimodal Integration in MagicHood

As illustrated in Figure 6.2 the implementation of our proposed mathematical model can be broadly divided into 3 modules, (i) **Pre-Processing module** analyses the input text and complex words are segregated. The rest of the text (referred as binding text) is translated to native language, (ii) **Explication module** processes the complex concepts and simplifies them into semantic elementary concepts and (iii) **Visualisation module** processes the simplified constructs and displays the final output into clickable multimodal format on a digital interface.

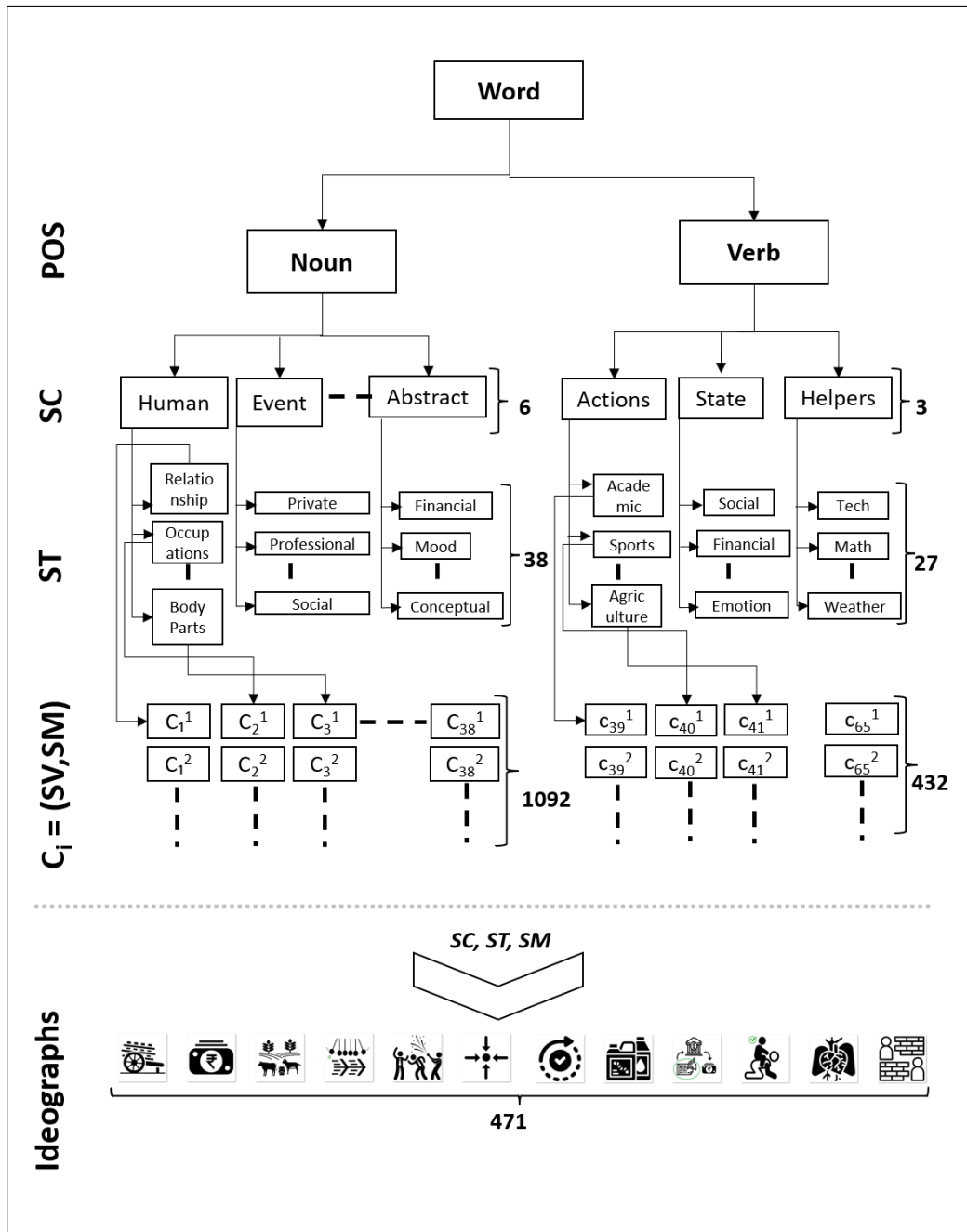
#### Pre-Process

Our experimental design starts with text pre-processing procedures which consumes the input text, and transforms the same into a format suitable for analysis in next phases. These transformations include, conversion to lowercase, tokenization, isolation of some characters (like punctuation, numeric entities, special characters etc.), Part-of-Speech (POS) tagging and lemmatization. Furthermore, for all nouns and verbs (identified via POS tags) we conduct an analysis of comprehension complexity using a variety of readability scores as outlined in the methodology. Through this analysis, we identify Complex-Words (CW), for ideographic conversion in subsequent stages of the process. The rest of the text (apart from CW) referred as "binding text" is translated to local/native language for the ease of end user readability. The arrangement/syntax of CW in the translated sentence may also suffer a change due to language related behavior etc. which is considered and re-aligned properly as needed.

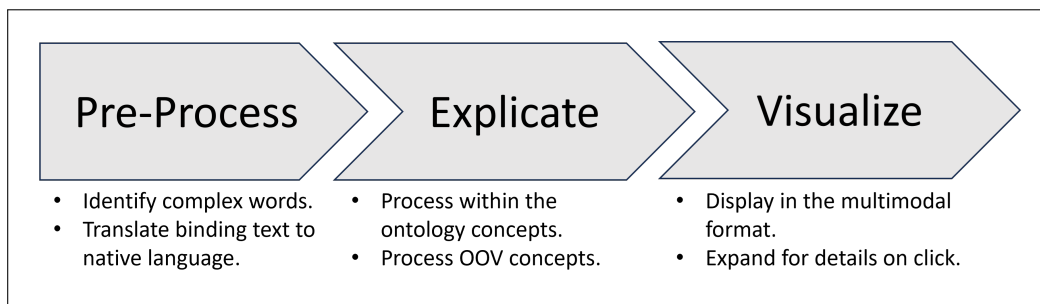
As an example, consider the following sentence.

Input Sentence: "I am going to mandi on motorcycle to buy seeds"

Complex-Words identified - CW1:mandi, CW2:motorcycle, CW3:seeds.



**Figure 6.1:** Hierarchical structure of our ontology.



**Figure 6.2:** High level process on experimental design.

Assuming the end user is native Marathi, and hence the binding text is translated to marathi for this example, the pre-processed sentence becomes:

मी <CW3> घेण्यासाठी <CW2> वर <CW1> जात आहे

## Explicate

In this module of our proposed approach, each complex word identified is broken down into hierarchy of semantic universals, including semantic class (SC), semantic template (ST) and explications of semantic variable (SV)-Semantic molecule (SM) tuples. There can be two scenarios while doing this conversion,

- The complex words identified are **already present in our ontology**. In such scenarios it becomes easy to retrieve the semantic structure, as this is just a search operation from knowledge base.
- The complex words identified are out of vocabulary and are **not present in the initial ontology**. In this scenario we need to leverage GenAI for explications.

In the example illustrated, CW2 and CW3 are present in the ontology, while CW1 is Out Of Vocabulary (OOV). Hence for CW2 and CW3, we refer ontology and break them down into semantic universals as below.

*CW2 - motorcycle*

*SC : Things*

*ST : Automobile*

*(SV, SM) tuples: (Category, Private Transport) (Wheels, Two)*

*CW3 - seeds*

*SC : Things*

*ST : Agro*

*(SV, SM) tuples: (Category, Germinate)*

For complex words not in ontology ("Mandi"/CW1 in this example), we use a sequence of prompts using few short learning, where relevant parts of the pre-compiled ontology is passed as context to a LLM and get the recommended response. The proposed method is similar to how a human annotator shall accomplish the task, taking cues from already described examples. The response from the LLM is parsed using output parsing methods in order to get the final output in expected template.

It is observed that LLM models exhibit good results in inferencing tasks, and our results are on the same lines [149]. We use Tree Of Thoughts (TOT) along with conversational chain and json response parser as the prompting technique in our

work. We illustrate the effectiveness empirically in results section. For the example illustrated earlier, where CW1("mandi") was OOV and to be explicated, the prompts are illustrated below.

### Prompt Engineering

Input Prompt (P1) - "Imagine the human annotator has been given the task to hierarchically break down a word into 2 levels. Level 1 considers of broad category of word and is referred as SC, Level 2 considers of narrow category of word and is referred as ST. Now considering the examples as illustrated *«Examples come here»*, use your inferencing to find the SC and ST for the word *«Word»*"

We use response parsing (json parser) to get the response in structured format. We use GPT 3.5 as the LLM to run this prompt, and we receive the following response.

Output json = "SC":"Location","ST":"Commercial"

Once the Semantic Template (ST) is identified ("Commercial" in this case), we share multiple examples of other concepts in the same ST, within the prompt and leverage LLM to break this down into SV, SM tuples as illustrated below.

Input Prompt (P2) - "Imagine the human annotator has been given the task to explain the semantic meaning of the word *«CW1»*, using Key-Value pairs. Keys to be considered should be from the predefined set of values *«add all SV elements for the ST "Commercial" »*. Values considers should be of a predefined set of values *«add all SM elements for the ST "Commercial"»*. Each semantic variable can only take limited values from semantic molecule sets as illustrated *«For "Commercial" ST add all semantic variable and molecule combinations here»*. Now considering the examples and constraints as illustrated use your inferencing to find the Key-Value pairs for the word *«Word»*. When there are multiple Key-Value pairs, use Key1,Value1, Key2,Value2 in your response."

When we run this for word "MANDI" and ST as "Commercial", we get the following response

Output json = "Key1":"Purpose","Value1":"Business"

Collating all the outputs from LLM, the final hierarchical semantic breakdown for the out of vocabulary concept *mandi* is illustrated below.

*CW1 - Mandi*

*SC : Location*

*ST : Commercial*

*(SV, SM) tuples: (Purpose, Business)*

**Table 6.3: Step by step illustrative message conversion.**

<p><b>Text Message (input) in English language for a Marathi end user:</b> I am going to mandi on motorcycle to buy seeds</p>
<p><b>Complex Word identification and Translation (to Marathi):</b> मी &lt;CW3&gt; घेण्यासाठी &lt;CW2&gt; वर &lt;CW1&gt; जात आहे</p>
<p><b>Entailment of complex words into hierarchical semantic universals:</b> &lt;entailment &gt; -&lt;cw &gt;seeds &lt;/ cw &gt; — &lt;sc &gt;things &lt;/ sc &gt; — &lt;st &gt;agro &lt;/ st &gt; —— &lt;sv &gt;category &lt;/ sv &gt;&lt;sm &gt;germinate &lt;/ sm &gt; - &lt;cw &gt;motorcycle &lt;/ cw &gt; — &lt;sc &gt;things &lt;/ sc &gt; — &lt;st &gt;automobile &lt;/ st &gt; —— &lt;sv &gt;category &lt;/ sv &gt;&lt;sm &gt;private transport &lt;/ sm &gt; —— &lt;sv &gt;wheels &lt;/ sv &gt;&lt;sm &gt;two &lt;/ sm &gt; - &lt;cw &gt;mandi &lt;/ cw &gt; — &lt;sc &gt;location &lt;/ sc &gt; — &lt;st &gt;commercial &lt;/ st &gt; —— &lt;sv &gt;purpose &lt;/ sv &gt;&lt;sm &gt;business &lt;/ sm &gt; &lt;/ entailment &gt;</p>

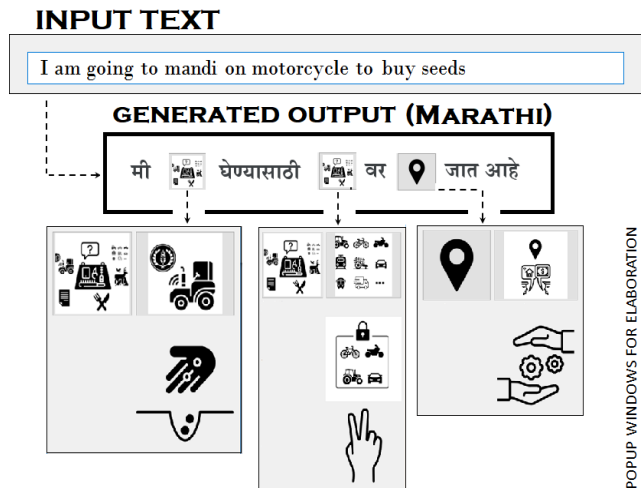
Once a new concept is explicated using the LLM, we add this to our ontology for easy reference in case the same word is encountered again in future. Table 6.3 illustrates this example with all hierarchical breakdown into multiple levels clearly.

## Visualize

While the inputs text is already broken down into semantic hierarchy of elementary concepts, as proposed in our mathematical model it is important to display the message using the power of digital platform into an user friendly and engageable manner. The semantic universals for complex words, including Semantic Class (SC), Semantic Template (ST) and Semantic Molecules (SM) (We don't display Semantic variables for ease of end users and the same was also acknowledged by our subjects in the experiment), are converted to a precompiled ideographic symbol.

We leverage *The Noun Project (NP)*<sup>2</sup> for the selection of appropriate ideograph for each SC, ST and SM in our work. NP is a crowd-sourced collection of 3 million icons created by designers from 120+ countries. Each element within semantic class, semantic template and semantic molecule maps to an unique ideograph. We have

<sup>2</sup><https://thenounproject.com/icons/>

**Final multimodal message output displayed to end user:****Figure 6.3: Final illustrative message (continued from Table 6.3)**

done an extensive exercise for the choice of appropriate ideograph, using participants from "digitally-neglected" category of semi-literates and taken majority voting [150].

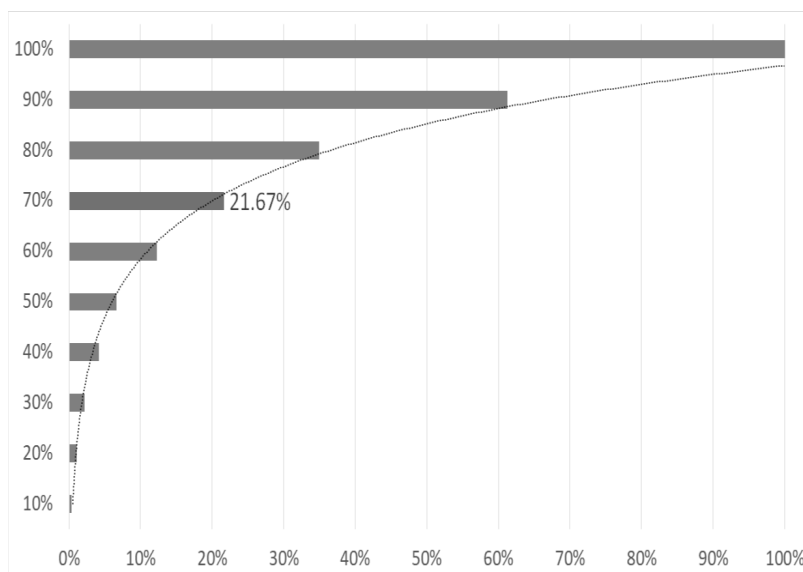
We illustrate the end to end example in Figure 6.3 for our example. The boxes in the bottom half appear only when the master icons (ideographs representing semantic classes) are clicked to make the entire experience more engaging.

## 6.5 Evaluating the Effectiveness of Proposed Magic-Hood Model: Empirical Findings and Discussions

We validate the proposed model on qualitative and quantitative parameters for a more holistic and nuanced assessment on performance and usefulness to end users. Having both parameters in our evaluation ensures that the proposed model not only meet technical benchmarks but also fulfill the practical requirements and expectations from digitally-neglected category of semi-literates in real-world settings.

We recruit 200 participants (123 males, 77 females; digitally-neglected semi-literates). We conduct the validation exercise over a five day window, where each day every participant interprets a set of multimodal sentences. The participants provide the interpretation, along with the feedback on experience and usability. At the end of the testing window, there is a final survey on ideographic effectiveness which every participant responds to judiciously. We have taken consent from all participants and did not collect any Personally Identifiable Information (PII) data in the process.





**Figure 6.4:** Ideographic volume following logarithmic growth with exploration of dataset. (x-axis is corpus size, and y-axis is ideographic count)

## Quantitative Validation

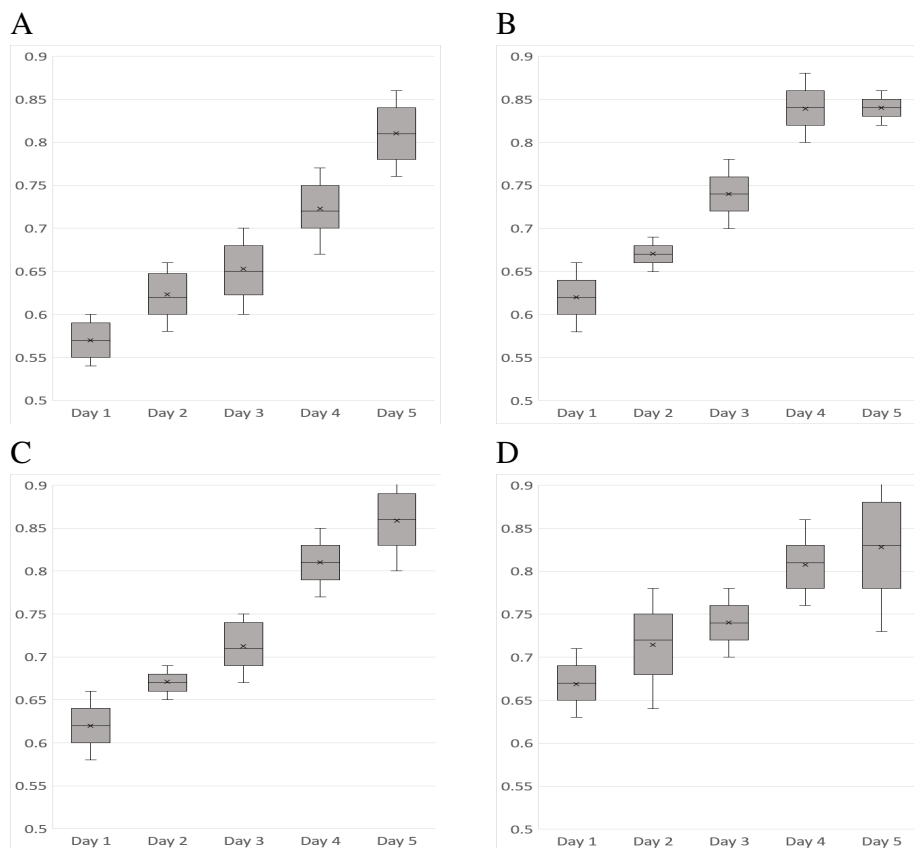
### Size of ideographic library

As illustrated in Figure 6.4 it is noteworthy that the total volume of semantic universals (which is linked to the number of ideographs) follows a logarithmic growth pattern and stabilizes as the dataset size increases. In the initial exploration of dataset, the rate of addition of new ideographs is high and the growth is noticeable however it becomes stable as the size reaches an equilibrium. Quantitatively, it is observed that while parsing the initial 21.67% of the corpus we identify the 70% of the ideographs. This analysis confirms that the size of ideographs will not become too big, and create learnability issues which was the case with earlier methods. This observation also aligns with the linguistic theory of Natural Semantic Metalanguage (NSM).

### Comprehensibility

We identify fifty sentences from our initial text dataset (described in Chapter 2 in detail) selected stochastically without replacement and converted these fifty sentences into multimodal output. We break these 50 sentences into groups of 10 each, so that over a 5 day period we can share those with our participants iteratively. The participants interpret the sentences and share their interpretation on a daily basis. We also share with them the right response at the end of the day. As illustrated in table 6.4 we evaluate the participant response on four semantic textual similarity metrics (i) METEOR [151], (ii) S-BERT [152], (iii) MPNet [153] and (iv) all-MiniLM-L6

**Table 6.4:** Empirical Evaluation (on 0-1 scale) on A: Meteor, B: S-BERT, C: MPNet and D: mini-LM.



[154]).

We derive two inferences from this exercise, firstly that the interpretation improves as participants become familiar with the process and secondly all the metrics indicate an effectiveness of 80% in semantic comprehensibility.

## Accuracy of OOV entailments

As illustrated in Table 6.5, we have applied a variety of prompt engineering techniques and observe the best results using Tree Of Thoughts (TOT) prompting methodology. We use GPT 3.5 as the choice of LLM since it is the most stable and acceptable LLM in the industry today. We explicate a set of 200 concepts (150 nouns and 50 verbs) randomly selected from our ontology (and not use these selected samples in prompt design), and generate semantic classes, semantic templates along with semantic variables and molecules tuples. We validate the output received from LLM with the human interpreted explications in our ontology and empirically analyse the results. We observe accuracy of more than 90% on all levels (SC, ST, SV, SM) using this approach. It is to be noted that once the OOV concepts are explicated, they are added to the master ontology for quick reference in future and avoid generative

**Table 6.5:** Empirical evaluation using multiple prompt engineering techniques.

	SC	ST	SV,SM
Chain of thought (COT)	0.88	0.92	0.86
Self-consistency COT	0.93	0.94	0.89
<b>Tree of Thoughts (TOT)</b>	<b>0.96</b>	<b>0.96</b>	<b>0.90</b>
Retrieval Augmented Generation (RAG)	0.94	0.96	0.89

entailments which can be costly and time consuming.

## Qualitative Validation

We perform quantitative validation on multiple aspects to validate our method.

## User Journey

The same participants, during the 5 day validation window also evaluate the system qualitatively via surveys on four parameters namely, (a) Expressiveness: ability to conceptualize the semantic characteristics of the inherent concept, (b) User Experience (UX): ease and convenience for users to interact and understand, (c) Intention to Reuse: measure of user satisfaction and perceived usefulness and (d) Interest: Measure the value, relevance and appeal to the audience. Table 6.6 illustrates the results from our survey, and it is a clear observation that participants have encountered a good experience on all parameters.

## Ideographic effectiveness

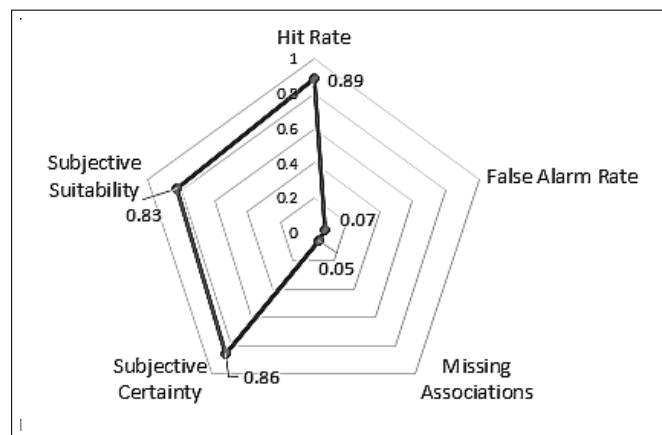
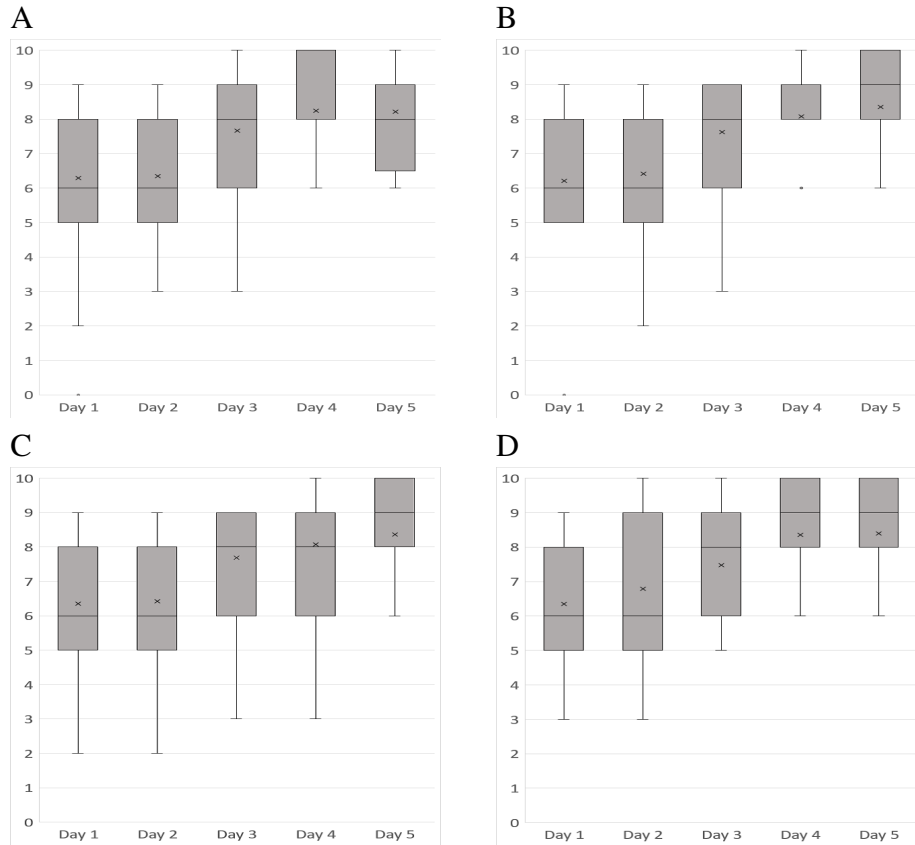
In order to evaluate the right choice of ideographs empirically we make use of Multiple Index Approach (MIA) as prescribed by European Telecommunications Standard Institute (ETSI). ETSI is designed on the principles of CCITT and ISO [155]. MIA recommends to evaluate ideographic effectiveness on 5 parameters namely, (a) Valid associations (hit-rate) (b) Invalid associations (false alarm rate) (c) Missing associations (d) Confidence of association (Subjective certainty) (e) Relevance of association (Subjective suitability).

As mentioned earlier, at the last day of our validation exercise we ask our participants to answer a final survey on these parameters where a set of ideographs are given to our participants and they match the concepts with available options, along with giving scores on the suitability and certainty of ideographs. We illustrate the results from this survey in Figure 6.5 normalised on (0-1) scale.

## Comparison with existing ideographic approaches

Appendix D illustrates some of the existing methods on ideographic communication (both linguistic and computational) as examples. While there is no direct one to

**Table 6.6:** Empirical Evaluation (on 0-10 scale) on A: Expressiveness, B: User Experience, C: Intention to Reuse and D: Interest.



**Figure 6.5:** Results for ideographic effectiveness using MIA

one comparison between our method, and the existing method there are a few points to highlight.

- **Learnability:** While existing ideographic communication methods contain thousands of ideographs (to the extent that PCS has 37,000 icons and BETA/ARASAAC has 12,000 icons), Magichood cleverly reduced this count to less than 500 using hierarchical framework and improves by 75x, thus increasing learnability.
- **Extensibility:** While existing communication methods are built for a particular domain or user base, Magichood is universal. The use of Generative-AI makes this universal in use.
- **Comprehensibility:** Pure ideographic scripts makes comprehensibility ambiguous, due to inherent native language of end users. Magichood, uses binding text to solve this and reduces ambiguity significantly.

## 6.6 Conclusion and Limitations

We propose a pioneering universal and innovative approach incorporating a versatile and cross-cultural multimodal ideographic metalanguage, alongside multilingual explanatory text, and elevate digital literacy for digitally-neglected semi-literate populations. We integrate AI techniques (NLP, ML, and GenAI) with the principles of the linguistic theory of NSM, and design a multimodal interactive communication method suitable for digital platforms. Our user interface (UI) design prioritizes fluidity and intuitiveness, streamlining navigation and minimizing potential sources of confusion. Proposed system addresses the gaps in existing methods and aligns with the objectives identified during our initial research phase. We validate the effectiveness of our methodology on multiple metrics and achieve good response. Furthermore, our design is adaptable for incorporating new concepts and seamlessly integrating additional domains and languages.

Our choice of ideographs is based on the selected options from Indian demography, and it can be made more universal with a more diverse and global participant base in future.



## **Part 5**

### **The Next Steps**





## 7 | Conclusions and Future Directions

*“The more we explore, the more we realize how much there is left to discover. The pursuit of knowledge is endless, and the journey is its own reward.”*

---

Marie Curie

Our endeavor is a stepping stone, to build a set of accessible and engageable socio-linguistic AI-models surrounding the digital enablement of semi-literates communities through the design and implementation of innovative methodologies and frameworks. We empathise, understand, and carefully propose AI methods for each category of semi-literates (digitally-naive, digitally-niche and digitally-neglected) independently, and not have one-size-fits-all methods. Leveraging state-of-the-art AI methods including Generative-AI, pretrained transformers along with algorithms from text and image analytics, we have addressed critical impediments hindering access to digital communication among all types of semi-literate populations quiet reasonably.

Our design methods focus on each category of semi-literates separately, and propose the AI-models which resolve the problems of end users, make it relevant for their practical use. We explore a variety of design principles and methods, in our work including the balance of the use of linguistic theories and computational linguistics algorithms, resulting in accumulation of datasets, pre-trained models, workable software all of which can be put to use right out of the box.

### **Technological Contributions**

We have made significant technological contributions in our work across multiple dimensions, which can be leveraged for further research by fellow researchers. Our work has been acknowledged and appreciated in multiple academic forums.

#### **1. Re-usable Datasets:**

As a part of our research, we have created re-usable datasets for fellow researchers.

- Our semi-literate text-message dataset has been created through extensive fieldwork, and contains 3,300 text messages and can be very useful for linguists and computer scientists.
  - We have created bilingual corpus for high resource languages and low resource languages, which can be leveraged for NLP models.
  - Our labelled image corpus on Takri serves as the largest available corpus for this language.
2. **OCR for no-data scripts:** VOLTAGE serves as a breakthrough to digitize languages which are not on digital ecosystem.
  3. **Alternate communication method using ideographs:** MagicHood is a stepping stone for taking "Digital India" to people who have bare-minimum exposure to academic literacy.
  4. **High-resource/low-resource bilingual data corpus:** Data is of prime importance to fuel training of AI models. Our unique method to build bilingual dataset without human intervention can help all remote languages worldwide solve problem with dataset availability.
  5. **A simple textual simplification paraphraser using data from closed captions:** Our method for textual simplification for semi-literates, using the creative use of data sourced from movies designed for children is a very simple but effective method for textual simplification.

Despite the notable achievements established in our work, it is imperative to acknowledge the dynamic and evolving nature of the AI technology and landscape, necessitating ongoing efforts for continuous improvement. While our methodologies have demonstrated efficacy in mitigating digital challenges and barriers for digital inclusion, the complex interplay of socio-economic, cultural, and infrastructural factors demands continuous exploration and refinement of the approaches.

Looking ahead, the pursuit of digital equality requires a steadfast commitment to innovation and collaboration across multidisciplinary domains. By harnessing emerging technologies such as Artificial Intelligence (AI), Machine Learning (ML), cloud computing, and Internet of Things (IoT), we can unlock new possibilities for enhancing the accessibility, relevance, and sustainability of digital interventions tailored to the needs of marginalized communities.

Moreover, our academic endeavour underscores the importance of robust evaluation frameworks and metrics to assess the impact and efficacy of digital enablement initiatives. Through rigorous empirical analysis, data-driven insights, taking periodic inputs and feedback from targeted audience, and leveraging the evolving AI technologies, the proposed methods can be iteratively refined and made more precise.

In conclusion, while significant accomplishment has been made in advancing the digital inclusion agenda, the journey towards accessible and engageable participation in the digital ecosystem is far from complete. It is incumbent upon researchers, practitioners, policymakers, and stakeholders to redouble their efforts, embrace technological innovation, and foster inclusive digital ecosystems that empower marginalized communities to thrive in the digital age.

We carve out short term achievable objectives in the immediate future and long term manageable tasks to facilitate incremental progress. While short term goals helps with the validation of research approach and refinement of methodologies along with iterative refinement with collaboration, long term goals help with continuity of research providing a vision for the future direction, potential expansion or further investigation, and a sense of purpose and direction.

## Short term goals

In the near future we plan to -

- **Maximizing reach via scaling language inclusion**

We shall use the proposed and well established AI-models in our work, to scale for multiple Indian languages and scripts. This includes training for OCR models using VOLTAGE as described in Chapter 4 and creation of bilingual corpus using content based image retrieval (CBIR) model as outlined in Chapter 5.

- **Grassroots outreach and adoption alliances**

While the AI-models have proven themselves in a smaller setup, we will partner with support groups who can contribute with taking these models to remote geographies, thereby helping to create an inclusive, empowered, and digitally literate societies. We will also seek help from government for access to data in museums and libraries, in order to digitize historical scripts and languages which are currently restricted and not accessible to majority of people.

## Long term goals

In the long term we will work towards accomplishment of two projects -

- **Preservation of linguistic and folk heritage**


Build an all inclusive technology ecosystem for every community, every language and script preserving and promoting access to documentary heritage, including digital and non-digital resources such as archives, libraries, and cultural artifacts. We plan to create an linguistic encyclopedia on Indian dialects and identify support levels for marginalized scripts and dialects, and work towards its digital inclusion.

- **"Connect every community" mission**

Enabling India to become a digitally empowered society and knowledge economy by promoting electronic e-governance which will happen only with the digital enablement of every citizen. We will work with like-minded people and groups to empower all categories of digitally semi-literate communities and onboard them to digital ecosystem. This will also empower all the citizens independent of their educational, societal, cultural and linguistic diversities to get enabled to participate in digital initiatives, access digital resources, engage with government and other stakeholders through digital platforms and make smarter and informed decisions.

## **8 | Appendices**

## Appendix A - Survey forms used for text message data collection



**BITS Pilani**  
Pilani Campus  
Department of Computer Science & Information Systems

पूरा नाम: \_\_\_\_\_

लिंग: पुरुष  महिला

उम्र: \_\_\_\_\_

गृहनगर: (शहर / राज्य) \_\_\_\_\_

शिक्षा: कभी स्कूल नहीं गया  5 वीं  8 वीं  10 वीं  12 वीं  स्नातक

क्या आप स्मार्टफोन का उपयोग करते हैं: हां  नहीं

क्या आप फोन मेसेज भेजना जानते हैं? हां  नहीं

आप कितनी बार फोन मेसेज करते हैं:

प्रत्येक दिन  सप्ताह में एक बार  महीने में एक बार  कभी नहीं

आप किसे फोन मेसेज भेजते हैं: परिवार  मित्र  अन्य \_\_\_\_\_

क्या आप अपने राज्य के बाहर अन्य क्षेत्रों के लोगों को संदेश भेजते हैं। हां  नहीं

आपका पेशा क्या है कृषि  दुकान  नौकरी  बेरोजगार  \_\_\_\_\_

स्थान महानगर  शहर  गांव

आपके द्वारा भेजे जाने वाले सबसे महत्वपूर्ण 10 फोन मेसेज। यदि कोई फोन मेसेज का आदान-प्रदान नहीं किया जाता है तो आप वॉयस कॉल उद्देश्य को संक्षेप में प्रस्तुत करेंगे।  
(फोवॉर्ड्स मेसेज, गुड मोर्निंग जैसे संदेशों का वर्णन न करें)

संदेश 1: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

संदेश 2: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

संदेश 3: भाषा \_\_\_\_\_

\_\_\_\_\_


\_\_\_\_\_

संदेश 4: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

**Figure 8.1:** Survey form for data collection - Page 1



**BITS Pilani**  
Pilani Campus  
Department of Computer Science & Information Systems

संदेश 5: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

संदेश 6: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

संदेश 7: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

संदेश 8: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

संदेश 9: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

संदेश 10: भाषा \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

इस फॉर्म की सारी जानकारी सही है। मैं समझता हूँ कि दी गई सभी सूचनाओं का उपयोग केवल अनुसंधान और लोक कल्याण के लिए किया जाएगा। सर्वेक्षण का यह उद्देश्य प्रवाल द्वारा पीएचडी के लिए डेटा एकत्र करना है जो अर्ध-साक्षर के लिए एनएलपी पर शोध कर रहे हैं।

दिनांक और हस्ताक्षर:

**Figure 8.2:** Survey form for data collection - Page 2

## Appendix B - Mathematical formulas of reading indices

$$\begin{aligned} \text{Flesch Reading Ease} &= 206.835 \\ &\quad - 1.015 \left( \frac{\text{Total Words}}{\text{Total Sentences}} \right) \\ &\quad - 84.6 \left( \frac{\text{Total Syllables}}{\text{Total Words}} \right) \end{aligned}$$

$$\begin{aligned} \text{Gunning Fog Index} &= 0.4 \left( \frac{\text{Total Words}}{\text{Total Sentences}} \right) \\ &\quad + 100 \left( \frac{\text{Complex Words}}{\text{Total Words}} \right) \end{aligned}$$

$$\begin{aligned} \text{Dale Chall Score} &= 0.1579 \left( \frac{\text{Difficult Words}}{\text{Total Words}} \times 100 \right) \\ &\quad + 0.0496 \left( \frac{\text{Total Words}}{\text{Total Sentences}} \right) \end{aligned}$$

$$\text{Adjusted Score} = \begin{cases} \text{Raw Score} & \text{if } \left( \frac{\text{Difficult Words}}{\text{Total Words}} \times 100 \right) \leq 5 \\ \text{Raw Score} + 3.6365 & \text{if } \left( \frac{\text{Difficult Words}}{\text{Total Words}} \times 100 \right) > 5 \end{cases}$$

$$\begin{aligned} \text{Flesch-Kincaid Grade Level} &= 0.39 \left( \frac{\text{Total Words}}{\text{Total Sentences}} \right) \\ &\quad + 11.8 \left( \frac{\text{Total Syllables}}{\text{Total Words}} \right) \\ &\quad - 15.59 \end{aligned}$$

$$\begin{aligned} \text{Coleman-Liau Index} &= 0.0588L - 0.296S - 15.8 \\ L &= \frac{\text{Total Letters}}{\text{Total Words}} \times 100 \\ S &= \frac{\text{Total Sentences}}{\text{Total Words}} \times 100 \end{aligned}$$



## Appendix C - Glyph feature store for VOLTAGE

Feature ID	Feature Description	Type	Range
F1	<b>Presence of Headline:</b> Checks for the presence of a horizontal line on the top of the sub-symbol.	B	0/1
F2	<b>Number of Loops:</b> Counts the number of loops in the symbol including loops with headline.	N	0-N
F3	<b>Number of Loops with headline:</b> Counts the number of loops the symbol makes with the headline (F1).	N	0-N
F4	<b>Presence of left-sidebar:</b> Check for the presence of a vertical line on the left-most side of sub-symbol.	B	0/1
F5	<b>Presence of right-sidebar:</b> Check for the presence of a vertical line on the right-most side of sub-symbol.	B	0/1
F6	<b>Number of connected components:</b> Counts the number of sub-symbols which are connected.	N	0-N
F7	<b>Number of endpoints:</b> Counts the number of points, which have only one black pixel in its 3x3 neighbourhood.	N	0-N
F8	<b>Number of junctions:</b> Counts the number of points, which have more than two black pixel in its 3x3 neighbourhood.	N	0-N
F9	<b>Number of junctions with headline:</b> Counts the number of junctions (F8) which touch the headline (F1).	N	0-N
F10	<b>Number of bend points clockwise:</b> Counts the number of points which makes 90 degree turn towards right direction.	N	0-N
F11	<b>Number of bend points anti-clockwise:</b> Counts the number of points which makes 90 degree turn towards left direction.	N	0-N
F12	<b>Aspect Ratio:</b> Ratio of symbol height and width on 0-100 scale.	N	0-100
F13	<b>Horizontal Symmetry:</b> Flip the image on Y axis (mirror the image in left-right perspective). Find the similarity with original image using threshold value.	B	0/1
F14	<b>Vertical Symmetry:</b> Flip the image on X axis (mirror the image in top-down perspective). Find the similarity with original image using threshold value.	B	0/1

F15	<b>Number of Dots:</b> Counts the number of points, which have zero black pixels in its 3x3 neighbourhood.	N	0-N
F16	<b>Number of left-right layers:</b> Counts the number of layers of black pixels on x axis. This relates to the maximum isolated black pixels in horizontal cross section.	N	0-N
F17	<b>Number of top-down layers:</b> Counts the number of layers of black pixels on y axis. This relates to the maximum isolated black pixels in vertical cross section.	N	0-N
F18	<b>Minimum horizontal projection:</b> Compute the horizontal projection (count the number of black pixels across y axis for every x axis) and find the minimum value. Scale this by taking ratio with height of image, and multiple by 100,	N	0-100
F19	<b>Minimum vertical projection:</b> Compute the vertical projection (count the number of black pixels across x axis for every y axis) and find the minimum value. Scale this by taking ratio with width of image, and multiple by 100,	N	0-100
F20	<b>Maximum horizontal projection:</b> Compute the horizontal projection (count the number of black pixels across y axis for every x axis) and find the maximum value. Scale this by taking ratio with height of image, and multiple by 100,	N	0-100
F21	<b>Maximum vertical projection:</b> Compute the vertical projection (count the number of black pixels across x axis for every y axis) and find the maximum value. Scale this by taking ratio with width of image, and multiple by 100,	N	0-100
F22	<b>Maximum left depth:</b> The maximum depth of the left profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100
F23	<b>Maximum right depth:</b> The maximum depth of the right profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100
F24	<b>Maximum top depth:</b> The maximum depth of the top profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100

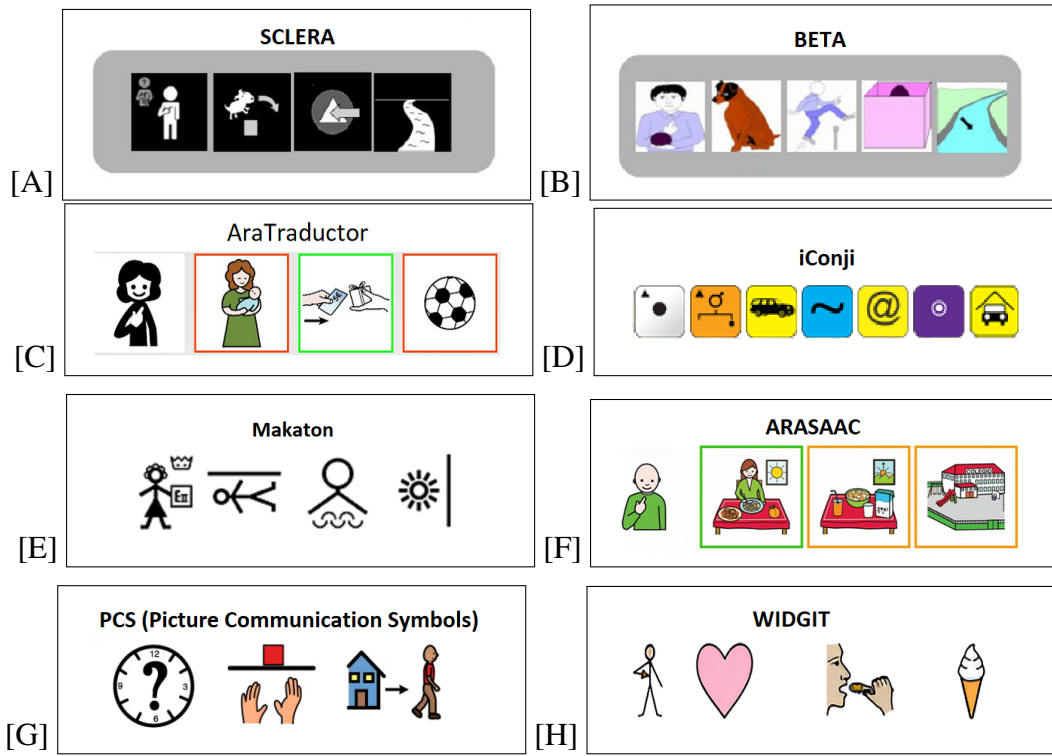
F25	<b>Maximum bottom depth:</b> The maximum depth of the bottom profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100
F26	<b>Minimum left depth:</b> The minimum depth of the left profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100
F27	<b>Minimum right depth:</b> The minimum depth of the right profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100
F28	<b>Minimum top depth:</b> The minimum depth of the top profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100
F29	<b>Minimum bottom depth:</b> The minimum depth of the bottom profile calculated as a percentage with respect to total width of the box enclosing the symbol.	N	0-100
F30	<b>Stroke length:</b> Count of total black pixels as a percentage with total area (height * width) of the symbol box.	N	0-100
F31	<b>Epicenter top down:</b> Find the mean of all black pixels, and compute the location of Y axis from center as a percentage from mid point on y axis .	N	0-100
F32	<b>Epicenter left right:</b> Find the mean of all black pixels, and compute the location of X axis from center as a percentage from mid point on x axis .	N	0-100

### Type B- Boolean, N - Number

#### Recommended Feature set for languages:

- Takri - F1, F2, F5, F7, F8, F12, F13, F14, F15.
- Modi - F1, F2, F4, F5, F7, F9, F12, F15, F16, F23, F30 .
- Ol Chiki - F2, F4, F8, F12, F16.
- Gujarati - F1, F2, F4, F5, F7, F12, F13, F15, F16.
- Wancho - F2, F5, F8, F12, F13, F15, F16, F21.

## Appendix D - Illustrative examples of existing visual methods of communication



	Method	Illustrated Example	Approximate ideograph count
A	SCLERA[51]	My dog jumped in the river.	3,500
B	BETA	My dog jumped in the river.	11,500
C	AraTradorctor[132]	My mum bought bought the soccer ball.	8,500
D	iConji[49]	My father's SUV is at the garage.	11,00
E	Makaton[156]	My queen died peacefully yesterday.	11,000
F	ARASAAC[157]	I eat breakfast at school.	12,000
G	PCS[158]	When d you want to go.	37,000
H	Widget	I like to eat ice cream.	15,000

## Bibliography

- [1] J. A. Van Dijk, “Digital divide research, achievements and shortcomings,” *Poetics*, vol. 34, no. 4-5, pp. 221–235, 2006.
- [2] J Van Dijk, *The deepening divide, inequality in the information society*. sage publications, 2005.
- [3] <https://ourworldindata.org/>, *Our world in data*, 2024.
- [4] <https://www.statista.com/>, *Statistica: Empowering people with data*, 2024.
- [5] <https://datacatalog.worldbank.org/>, *The world bank: Data catalogue*, 2024.
- [6] <https://www.gapminder.org/>, *Gapminder*, 2024.
- [7] <https://datareportal.com/global-digital-overview>, *Datareportal: Digital around the world*, 2024.
- [8] <https://www.pbl.nl/en/hyde-history-database-of-the-global-environment>, *History database of the global environment (hyde)*, 2024.
- [9] Kezang and J. Whalley, “Closing the digital divide: The role of services and infrastructure in bhutan,” *Prometheus*, vol. 25, no. 1, pp. 69–84, 2007.
- [10] V. M. Pitter, “Identifying core components of digital literacy initiatives for adult education programs,” Ph.D. dissertation, Walden University, 2022.
- [11] D. M. Eberhard, G. F. Simons, and C. D. Fenning, *Ethnologue: Languages of the world*, 2015.
- [12] R. L. Myoya, F. Banda, V. Marivate, and A. Modupe, “Fine-tuning multilingual pretrained african language models,” in *4th Workshop on African Natural Language Processing*, 2023.
- [13] <https://unesdoc.unesco.org/ark:/48223/pf0000192416>, *Unesco project: Atlas of the world’s languages in danger*, 2011.
- [14] I. Medhi, S. Patnaik, E. Brunskill, S. N. Gautama, W. Thies, and K. Toyama, “Designing mobile interfaces for novice and low-literacy users,” *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 18, no. 1, pp. 1–28, 2011.

- [15] I. Medhi, A. Sagar, and K. Toyama, “Text-free user interfaces for illiterate and semi-literate users,” in *2006 international conference on information and communication technologies and development*, IEEE, 2006, pp. 72–82.
- [16] T. Nayyar, S. Aggarwal, D. Khatter, K. Kumar, S. Goswami, and L. Saini, “Opportunities and challenges in digital literacy: Assessing the impact of digital literacy training for empowering urban poor women,”
- [17] <https://asercentre.org/>, *Annual status of education report*.
- [18] <https://ihds.umd.edu/>, *India human development survey*.
- [19] A. Wierzbicka, *Semantics: Primes and universals: Primes and universals*. Oxford University Press, UK, 1996.
- [20] C. Goddard, “Natural semantic metalanguage: The state of the art,” *Cross-linguistic semantics*, vol. 102, pp. 1–34, 2008.
- [21] M. Vila, M. A. Martí, and H. Rodríguez, “Paraphrase concept and typology. a linguistically based and computationally oriented approach,” *Procesamiento del lenguaje natural*, no. 46, pp. 83–90, 2011.
- [22] R. de Salvo Braz, R. Girju, V. Punyakanok, D. Roth, and M. Sammons, “An inference model for semantic entailment in natural language,” in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, Springer, 2006, pp. 261–286.
- [23] A. H. Al-Ghidani and A. A. Fahmy, “Conditional text paraphrasing: A survey and taxonomy,” *International journal of advanced computer science and applications*, vol. 9, no. 11, pp. 589–594, 2018.
- [24] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
- [25] P. Sharma and N. Goyal, “Zero-shot reductive paraphrasing for digitally semi-literate,” in *Forum for Information Retrieval Evaluation*, 2021, pp. 91–98.
- [26] A. Roy and D. Grangier, “Unsupervised paraphrasing without translation,” *arXiv preprint arXiv:1905.12752*, 2019.
- [27] G. Lample, A. Conneau, L. Denoyer, and M. Ranzato, “Unsupervised machine translation using monolingual corpora only,” *arXiv preprint arXiv:1711.00043*, 2017.

- [28] Z. Li, X. Jiang, L. Shang, and H. Li, "Paraphrase generation with deep reinforcement learning," *arXiv preprint arXiv:1711.00279*, 2017.
- [29] A. Aggarwal, M. Mittal, and G. Battineni, "Generative adversarial network: An overview of theory and applications," *International Journal of Information Management Data Insights*, vol. 1, no. 1, p. 100 004, 2021.
- [30] A. K. Singh, "Natural language processing for less privileged languages: Where do we come from? where are we going?" In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, 2008.
- [31] C. Cieri, M. Maxwell, S. Strassel, and J. Tracey, "Selection criteria for low resource language programs," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4543–4549.
- [32] P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury, "The state and fate of linguistic diversity and inclusion in the nlp world," *arXiv preprint arXiv:2004.09095*, 2020.
- [33] A. Chaudhuri *et al.*, *Optical character recognition systems*. Springer, 2017.
- [34] B. Chaudhuri, "On ocr of major indian scripts: Bangla and devanagari," in *Guide to OCR for Indic Scripts*, Springer, 2009, pp. 27–42.
- [35] R. Smith, "An overview of the tesseract ocr engine," in *Ninth international conference on document analysis and recognition (ICDAR 2007)*, IEEE, vol. 2, 2007, pp. 629–633.
- [36] T. M. Breuel, "The ocropus open source ocr system," in *Document recognition and retrieval XV*, SPIE, vol. 6815, 2008, pp. 120–134.
- [37] B. Kiessling, M. T. Miller, M. Romanov, and S. B. Savant, "Important new developments in arabographic optical character recognition (ocr)," *Al-Uşūr al-Wuṣṭā*, vol. 25, no. 1, p. 1, 2017.
- [38] C. Wick, C. Reul, and F. Puppe, "Calamari-a high-performance tensorflow-based deep learning package for optical character recognition," *arXiv preprint arXiv:1807.02004*, 2018.
- [39] M. Tomaschek, "Evaluation of off-the-shelf ocr technologies," *Bachelor thesis Masaryk University, Brno, Czech Republic*, 2018.
- [40] A. Ragni, K. M. Knill, S. P. Rath, and M. J. Gales, "Data augmentation for low resource languages," in *INTERSPEECH 2014: 15th Annual Conference of the International Speech Communication Association*, International Speech Communication Association (ISCA), 2014, pp. 810–814.

- 
- [41] S. Y. Feng *et al.*, “A survey of data augmentation approaches for nlp,” *arXiv preprint arXiv:2105.03075*, 2021.
- [42] B. Li, Y. Hou, and W. Che, “Data augmentation approaches in natural language processing: A survey,” *AI Open*, 2022.
- [43] H. Nawar, “Multicultural transposition: From alphabets to pictographs, towards semantographic communication,” *Technoetic Arts*, vol. 10, no. 1, pp. 59–68, 2012.
- [44] T. Takasaki, “Pictnet: Semantic infrastructure for pictogram communication,” in *The 3rd International WordNet Conference (GWC-06)*, Citeseer, 2006, pp. 279–284.
- [45] A. Michaud, “Pictographs and the language of naxi rituals,” *Arnoldsche Art Publishers*, 2011.
- [46] C. Burke, “Isotype representing social facts pictorially,” *Information Design Journal*, 2009.
- [47] L. A. Archer, “Blissymbolics—a nonverbal communication system,” *Journal of Speech and Hearing Disorders*, 1977.
- [48] M. Randic, *Nobel Universal Graphical Language*. Xlibris Corporation, 2010.
- [49] K. Tatti, “New iconji language for the symbol-minded-bizwest,” *BizWest*, 2016.
- [50] H. Saggion, E. Gómez-Martínez, E. Etayo, A. Anula, and L. Bourg, “Text simplification in simplext: Making texts more accessible,” *Procesamiento del lenguaje natural*, 2011.
- [51] L. Sevens, V. Vandeghinste, I. Schuurman, and F. Van Eynde, “Natural language generation from pictographs,” in *Proceedings of the 15th European Workshop on Natural Language Generation (ENLG)*, 2015, pp. 71–75.
- [52] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, “Improving language understanding by generative pre-training,” 2018.
- [53] H. Touvron *et al.*, “Llama: Open and efficient foundation language models,” *arXiv preprint arXiv:2302.13971*, 2023.
- [54] Y. Chang *et al.*, “A survey on evaluation of large language models,” *arXiv preprint arXiv:2307.03109*, 2023.
- [55] P. Sharma, N. Goyal, and M. Vinay, “Semi-literate texting (slt): Survey based text message dataset from digitally semi-literate users in india,” *Data in Brief*, 2021.



- [56] P. Gilster and P. Glister, *Digital literacy*. Wiley Computer Pub. New York, 1997.
- [57] J. Coldwell-Neilson, T. Cooper, and N. Patterson, “Capability demands of digital service innovation,” in *Leadership, Management, and Adoption Techniques for Digital Service Innovation*, IGI Global, 2020, pp. 45–64.
- [58] T. Nayyar, S. Aggarwal, D. Khatter, K. Kumar, S. Goswami, and L. Saini, “Opportunities and challenges in digital literacy: Assessing the impact of digital literacy training for empowering urban poor women,” *University of Delhi*, 2019.
- [59] A. S. Khokhar, “Digital literacy: How prepared is india to embrace it?” *International Journal of Digital Literacy and Digital Competence (IJDLC)*, vol. 7, no. 3, pp. 1–12, 2016.
- [60] <https://www.digitalindia.gov.in/>, *Digitalindia digital india programme | government of india*, 2020.
- [61] M. Burch, S. Lohmann, D. Pompe, and D. Weiskopf, “Prefix tag clouds,” in *2013 17th International Conference on Information Visualisation*, IEEE, 2013, pp. 45–50.
- [62] <https://cmsindia.org/>, *Center for media studies (cms)*, 2021.
- [63] R. Flesch, “How to write plain english: A book for lawers and consumers,” *HeinOnline*, 2014.
- [64] L. N. Kennette and N. A. Wilson, “Universal design for learning (udl): Student and faculty perceptions,” *Journal of Effective Teaching in Higher Education*, vol. 2, no. 1, pp. 1–26, 2019.
- [65] J. Kirkman, C. Snow, and I. Watson, “Controlled english as an alternative to multiple translations,” *IEEE transactions on professional communication*, no. 4, pp. 159–161, 1978.
- [66] A. SE-Guide, “Aecma simplified english,” *AECMA Document: PSC-85-16598*, AECMA, Bruxelles, 1995.
- [67] Y. Xian, B. Schiele, and Z. Akata, “Zero-shot learning-the good, the bad and the ugly,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4582–4591.
- [68] W. Wang, V. W. Zheng, H. Yu, and C. Miao, “A survey of zero-shot learning: Settings, methods, and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.

- [69] M. Johnson *et al.*, “Google’s multilingual neural machine translation system: Enabling zero-shot translation,” *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 339–351, 2017.
- [70] M. Kay, “The proper place of men and machines in language translation,” *machine translation*, vol. 12, no. 1-2, pp. 3–23, 1997.
- [71] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, “Multi-task learning for multiple language translation,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1723–1732.
- [72] O. Firat, K. Cho, and Y. Bengio, “Multi-way, multilingual neural machine translation with a shared attention mechanism,” *arXiv preprint arXiv:1601.01073*, 2016.
- [73] J. Mallinson, R. Sennrich, and M. Lapata, “Paraphrasing revisited with neural machine translation,” in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 2017, pp. 881–893.
- [74] B. Zoph and K. Knight, “Multi-source neural translation,” *arXiv preprint arXiv:1601.00710*, 2016.
- [75] O. Firat, B. Sankaran, Y. Al-Onaizan, F. T. Y. Vural, and K. Cho, “Zero-resource translation with multi-lingual neural machine translation,” *arXiv preprint arXiv:1606.04164*, 2016.
- [76] P. Lison and J. Tiedemann, “Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles,” 2016.
- [77] J. Tiedemann, “Improved sentence alignment for movie subtitles,” in *Proceedings of RANLP*, vol. 7, 2007.
- [78] J. Tiedemann and L. Nygaard, “The opus corpus-parallel and free: [Http://logos.uio.no/opus/](http://logos.uio.no/opus/),” in *LREC*, Citeseer, 2004.
- [79] A. M. Turk, “Amazon mechanical turk,” *Retrieved August*, vol. 17, p. 2012, 2012.
- [80] B. Sagot and D. Fišer, “Building a free french wordnet from multilingual resources,” *OntoLex*, 2008.
- [81] B. Hamp and H. Feldweg, “Germanet-a lexical-semantic net for german,” in *Automatic information extraction and building of lexical semantic resources for NLP applications*, 1997.

- [82] Y. J. Choe, K. Park, and D. Kim, “Word2word: A collection of bilingual lexicons for 3,564 language pairs,” *arXiv preprint arXiv:1911.12019*, 2019.
- [83] <https://wordnet.princeton.edu/>, *A lexical database for english*, 2010.
- [84] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, 2002, pp. 311–318.
- [85] W. Xu, C. Napoles, E. Pavlick, Q. Chen, and C. Callison-Burch, “Optimizing statistical machine translation for text simplification,” *Transactions of the Association for Computational Linguistics*, vol. 4, pp. 401–415, 2016.
- [86] E. Reiter, “A structured review of the validity of bleu,” *Computational Linguistics*, vol. 44, no. 3, pp. 393–401, 2018.
- [87] D. Kumar, L. Mou, L. Golab, and O. Vechtomova, “Iterative edit-based unsupervised sentence simplification,” *ACL*, 2020.
- [88] L. Martin, B. Sagot, E. de la Clergerie, and A. Bordes, “Controllable sentence simplification,” *arXiv preprint arXiv:1910.02677*, 2019.
- [89] E. Sulem, O. Abend, and A. Rappoport, “Simple and effective text simplification using semantic and neural methods,” *arXiv preprint arXiv:1810.05104*, 2018.
- [90] T. Vu, B. Hu, T. Munkhdalai, and H. Yu, “Sentence simplification with memory-augmented neural networks,” *arXiv preprint arXiv:1804.07445*, 2018.
- [91] X. Zhang and M. Lapata, “Sentence simplification with deep reinforcement learning,” *arXiv preprint arXiv:1703.10931*, 2017.
- [92] S. Zhao, R. Meng, D. He, S. Andi, and P. Bambang, “Integrating transformer and paraphrase rules for sentence simplification,” *arXiv:1810.11193*, 2018.
- [93] P. Sharma, P. Goyal, V. Sharma, and N. Goyal, “Voltage: A versatile contrastive learning based ocr methodology for ultra low-resource scripts through auto glyph feature extraction,” in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 881–899.
- [94] L. Torrey and J. Shavlik, “Transfer learning,” in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, IGI global, 2010, pp. 242–264.
- [95] F. Zhuang *et al.*, “A comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

- [96] G. A. Grierson, “Linguistic survey of india. western hindi and panjabi,” vol. 1, no. 9, 1918.
- [97] G. A. Grierson, “The languages of the northern himalayas, being studies in the grammar of twenty-six himalayan dialects. by the revt. grahame bailey, bd, ma, mras asiatic society monographs, vol. xii. london, 1908.,” *Journal of the Royal Asiatic Society*, vol. 41, no. 1, pp. 184–189, 1909.
- [98] P. T. Daniels, “Writing systems,” *The handbook of linguistics*, pp. 75–94, 2017.
- [99] L. Likforman-Sulem, A. Zahour, and B. Taconet, “Text line segmentation of historical documents: A survey,” *International Journal of Document Analysis and Recognition (IJ DAR)*, vol. 9, no. 2, pp. 123–138, 2007.
- [100] A. A. Shinde and D. Chougule, “Text pre-processing and text segmentation for ocr,” *International Journal of Computer Science Engineering and Technology*, vol. 2, no. 1, pp. 810–812, 2012.
- [101] S. Magotra, B. Kaushik, and A. Kaul, “Use of classification approaches for takri text challenges,” in *Information and Communication Technology for Competitive Strategies (ICTCS 2020)*, Springer, 2021, pp. 403–410.
- [102] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, and Y. Chao, “The connected-component labeling problem: A review of state-of-the-art algorithms,” *Pattern Recognition*, vol. 70, pp. 25–43, 2017.
- [103] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, and L. Van Gool, “Scan: Learning to classify images without labels,” in *European conference on computer vision*, Springer, 2020, pp. 268–285.
- [104] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and repeats?” *Pattern Recognition*, vol. 93, pp. 95–112, 2019.
- [105] S. Tripathi, S. Chandra, A. Agrawal, A. Tyagi, J. M. Rehg, and V. Chari, “Learning to generate synthetic data via compositing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 461–470.
- [106] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*, Springer, 2014, pp. 818–833.
- [107] R. C. Staudemeyer and E. R. Morris, “Understanding lstm—a tutorial into long short-term memory recurrent neural networks,” *arXiv preprint arXiv:1909.09586*, 2019.

- [108] P. Khosla *et al.*, “Supervised contrastive learning,” *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.
- [109] D. Lopresti, “Optical character recognition errors and their effects on natural language processing,” in *Proceedings of the second workshop on Analytics for Noisy Unstructured Text Data*, 2008, pp. 9–16.
- [110] R. Holley, “How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs,” *D-Lib Magazine*, vol. 15, no. 3/4, 2009.
- [111] P. Sharma, N. Goyal, and P. Goyal, “A fully automated and scalable parallel data augmentation for low resource languages using image and text analytics,” in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, 2023, pp. 358–361.
- [112] A. Magueresse, V. Carles, and E. Heetderks, “Low-resource languages: A review of past work and future challenges,” *arXiv preprint arXiv:2006.07264*, 2020.
- [113] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, and R. Kaur, “Neural machine translation for low-resource languages: A survey,” *ACM Computing Surveys*, vol. 55, no. 11, pp. 1–37, 2023.
- [114] A. Antonacopoulos, D. Bridson, C. Papadopoulos, and S. Pletschacher, “A realistic dataset for performance evaluation of document layout analysis,” in *2009 10th International Conference on Document Analysis and Recognition*, IEEE, 2009, pp. 296–300.
- [115] M. University of Salford, *Pattern recognition image analysis*, 2004.
- [116] G. Bradski and A. Kaehler, “Opencv,” *Dr. Dobb’s journal of software tools*, vol. 3, p. 2, 2000.
- [117] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [118] H. Bay, T. Tuytelaars, and L. Van Gool, “Surf: Speeded up robust features,” in *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, Springer, 2006, pp. 404–417.
- [119] GoogleAI, *Language-agnostic bert sentence embedding*, 2020.
- [120] P. F. Brown, J. C. Lai, and R. L. Mercer, “Aligning sentences in parallel corpora,” in *29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 169–176.

- [121] A. F. X. Maffei, *An English-Konkani Dictionary*. Basel Mission Press, 1883.
- [122] E. Agirre *et al.*, “Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation,” in *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.*, ACL (Association for Computational Linguistics), 2016.
- [123] G. Ramesh *et al.*, “Samanantar: The largest publicly available parallel corpora collection for 11 indic languages,” *arXiv preprint arXiv:2104.05596*, 2021.
- [124] J. H. Zar, “Spearman rank correlation,” *Encyclopedia of biostatistics*, vol. 7, 2005.
- [125] B. Zoph, D. Yuret, J. May, and K. Knight, “Transfer learning for low-resource neural machine translation,” *arXiv preprint arXiv:1604.02201*, 2016.
- [126] E.-S. A. Lee *et al.*, “Pre-trained multilingual sequence-to-sequence models: A hope for low-resource language translation?” *arXiv preprint arXiv:2203.08850*, 2022.
- [127] L. Xue *et al.*, “Mt5: A massively multilingual pre-trained text-to-text transformer,” *arXiv preprint arXiv:2010.11934*, 2020.
- [128] K. Revanuru, K. Turlapaty, and S. Rao, “Neural machine translation of indian languages,” in *Proceedings of the 10th annual ACM India compute conference*, 2017, pp. 11–20.
- [129] P. Sharma, N. Goyal, and P. Goyal, “Multimodal semantographic metalanguage (msm): A novel methodology for digital enablement of semi-literates,” in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, 2023, pp. 844–851.
- [130] A. Van Dam, “Computer software for graphics,” *Scientific American*, 1984.
- [131] A. G. Karkar, J. M. Alja’am, and A. Mahmood, “Illustrate it! an arabic multimedia text-to-picture m-learning system,” 2017.
- [132] S. Bautista, R. Hervás, A. Hernández-Gil, C. Martínez-Díaz, S. Pascua, and P. Gervás, “Aratrador: Text to pictogram translation using natural language processing techniques,” in *Proceedings of the XVIII International Conference on Human Computer Interaction*, 2017.
- [133] H. Li, J. Tang, G. Li, and T.-S. Chua, “Word2image: A system for visual interpretation of concepts,” *Internet Multimedia Search and Mining*, 2013.

- 
- [134] A. G. Karkar, “An ontology based text-to-picture multimedia m-learning system,” Ph.D. dissertation, 2018.
- [135] H. Yang, J. Lin, A. Yang, P. Wang, C. Zhou, and H. Yang, “Prompt tuning for generative multimodal pretrained models,” *arXiv preprint arXiv:2208.02532*, 2022.
- [136] Q. Sun *et al.*, “Generative pretraining in multimodality,” *arXiv:2307.05222*, 2023.
- [137] openAI, *Dall.e: creating images from text*, 2024.
- [138] Google, *Imagen 2*, 2024.
- [139] Stability.AI, *Generative ai using dream studio*, 2024.
- [140] A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell, “Itri-04-08 the sketch engine,” *Information Technology*, vol. 105, no. 116, pp. 105–116, 2004.
- [141] M. Kunilovskaya and M. Koviiazina, “Sketch engine: A toolbox for linguistic discovery,” *Jazykovedny Casopis*, vol. 68, no. 3, p. 503, 2017.
- [142] M. A. H. Taieb, T. Zesch, and M. B. Aouicha, “A survey of semantic relatedness evaluation datasets and procedures,” *Artificial Intelligence Review*, 2020.
- [143] X. Xue, H. Wang, J. Zhang, Y. Huang, M. Li, and H. Zhu, “Matching transportation ontologies with word2vec and alignment extraction algorithm,” *Journal of Advanced Transportation*, 2021.
- [144] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.
- [145] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih, “Dissecting contextual word embeddings: Architecture and representation,” *arXiv preprint arXiv:1808.08949*, 2018.
- [146] T. Zhang, R. Ramakrishnan, and M. Livny, “Birch: A new data clustering algorithm and its applications,” *Data Mining and Knowledge Discovery*, 1997.
- [147] L. D. Baker and A. K. McCallum, “Distributional clustering of words for text classification,” in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 96–103.
- [148] C. Comito, A. Forestiero, and C. Pizzuti, “Word embedding based clustering to detect topics in social media,” in *IEEE/WIC/ACM International Conference on Web Intelligence*, 2019, pp. 192–199.

- [149] Y. Wang, K. Chen, H. Tan, and K. Guo, “Tabi: An efficient multi-level inference system for large language models,” in *Proceedings of the Eighteenth European Conference on Computer Systems*, 2023, pp. 233–248.
- [150] <https://thenounproject.com/about/>, *Icons and photos for everything*, 2010.
- [151] S. Banerjee and A. Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005.
- [152] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.
- [153] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, “Mpnet: Masked and permuted pre-training for language understanding,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 857–16 867, 2020.
- [154] <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>, *Sentence-transformers*, 2022.
- [155] M. Böcker, “A multiple index approach for the evaluation of pictograms and icons,” *Computer Standards & Interfaces*, 1996.
- [156] I. Deliyannis, C. Simpsiri, and P. Tsirigoti, “Interactive multimedia learning for children with communication difficulties using the makaton method,” in *International Conference on Information Communication Technologies in Education*, 2008, pp. 10–12.
- [157] R. Hervás, S. Bautista, G. Méndez, P. Galván, and P. Gervás, “Predictive composition of pictogram messages for users with autism,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, pp. 5649–5664, 2020.
- [158] S. Dada, A. Huguet, and J. Bornman, “The iconicity of picture communication symbols for children with english additional language and mild intellectual disability,” *Augmentative and Alternative Communication*, vol. 29, no. 4, pp. 360–373, 2013.



## **Publications of the Candidate**

---

### **List of research article(s) in international conferences (published)**

1. **Prawaal**, Goyal, P., Sharma, V., & Goyal, N. (2024, April). VOLTAGE: A Versatile Contrastive Learning based OCR Methodology for ultra low-resource scripts through Auto Glyph Feature Extraction. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 881-899).
2. **Prawaal**, Goyal, N., & Goyal, P. (2023, March). Multimodal Semantographic Metalanguage (MSM): A novel methodology for digital enablement of semi-literates. In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing (pp. 844-851).
3. **Prawaal**, Goyal, N., & Goyal, P. (2023, March). A fully automated and scalable Parallel Data Augmentation for Low Resource Languages using image and text analytics. In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing (pp. 358-361).
4. **Prawaal**, & Goyal, N. (2021, December). Zero-shot reductive paraphrasing for digitally semi-literate. In Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation (pp. 91-98).

### **List of research article(s) in journal (published)**

5. **Prawaal**, Goyal N, Vinay MR. Semi-literate Texting (SLT): Survey based text message dataset from digitally semi-literate users in India. Data in brief. 2021 Oct 1;38:107329.

### **List of research article(s) (submitted)**

6. **Prawaal.**, Goyal, P., Sharma, V., & Goyal, N. MagicHood: Multimodal Generative-AI powered Ideographic Code, with Hierarchical Ontology for Digital platforms.

## Patents Filed

Patent Details	Status
VOLTAGE - An OCR methodology for low-to-no-data scenarios	Patent Applied (March 2024)
MagicHood - Generative AI powered multimodal metalanguage	Patent Applied (June 2024)

## Re-useable Datasets

About	Location
Semi-literate text message dataset (382 respondents; 3368 messages)	<a href="https://github.com/prawaal/Semi-Literate-Text">https://github.com/prawaal/Semi-Literate-Text</a>
Bilingual dataset on simplified vocabulary for German-English, French-English and Spanish-English combinations (each 270K approx.)	<a href="https://github.com/prawaal/ZSL-Paraphrasing/tree/main/Data">https://github.com/prawaal/ZSL-Paraphrasing/tree/main/Data</a>
Konkani-Marathi Bilingual corpus (15K mapped sentences)	<a href="https://github.com/prawaal/Konkani-Marathi-Data-Corpus">https://github.com/prawaal/Konkani-Marathi-Data-Corpus</a>
Labelled symbols for Takri (226K labelled symbols)	<a href="https://github.com/prawaal/Takri">https://github.com/prawaal/Takri</a>

## *Brief Biography of the Candidate*



**Prawaal**, a Principal Data Scientist at Infosys Pune, with over 24 years of IT experience is a part of Data & Advanced Analytics practice at the Pune campus. Over the last 24 years, Prawaal has played various roles including applied research, consulting, data & analytics project execution and has participated in multiple transformational projects across large multinational customers, harnessing data to unveil invaluable insights. In his current role works with large clients with ai-advisory to bridge the gap between business objectives and AI capabilities, enabling organizations to harness the full potential of AI to drive strategic growth and competitive advantage.

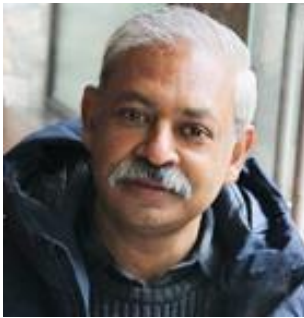
Prawaal earned his B.Tech. degree in Computer Science & Engineering from N.I.T., Hamirpur (formerly known as R.E.C.), located in Himachal Pradesh (H.P.), India, in the year 2000. After working in the industry for 15 years, he accomplished his M.Tech. in Software Systems specialising in Analytics from BITS Pilani, India, between 2015-2017. Thereafter he started his PhD journey with the broader area on ai-for-good enabling digital communication for marginalised communities in Indian demography.

Prawaal's expertise lies primarily in Natural Language Processing (NLP), with a particular emphasis on addressing challenges related to Low Resource Languages (LRL) and empowering semi-literate individuals to access and utilize information effectively. Over the years, Prawaal has made significant contributions to the field, publishing several research papers in esteemed journals and conferences.

Motivated by an insatiable curiosity and a deep-seated commitment to leveraging technology for societal advancement, Prawaal remains dedicated to exploring innovative solutions and pushing the boundaries of NLP beyond popular languages on the internet. Through his work, he endeavors to equip under-represented languages and individuals with technology aids, ultimately driving positive impact and more inclusive future for everyone.



## *Brief Biography of the Supervisor*



**Professor. Navneet Goyal**, a distinguished academician and researcher, completed his Ph.D. in Mathematics from the University of Roorkee (now IIT Roorkee) in 1995. Subsequently, he embarked on a journey in academia, joining BITS-Pilani in the same year, where he has remained a stalwart ever since. Presently, he holds the esteemed position of Senior Professor and Head of the Department of Computer Science & Information Systems at BITS-Pilani.

Throughout his illustrious career, Prof. Goyal has made significant contributions to the field of computer science, particularly in the realms of Big Data Analytics, Data Mining, and Machine Learning along with Natural Language Processing. His dedication and passion for research have led him to explore innovative avenues and build partnerships via MOUs with various universities worldwide.

Prof. Goyal's remarkable achievements have garnered widespread recognition, including the prestigious IBM Scalable Data Analytics Innovation Faculty Award in 2010. This award, bestowed under the Smarter Planet Initiative, honored his groundbreaking research concept titled "Developing a Smart Crop Management System using Data Analytics." Moreover, he has successfully led numerous funded projects, exemplified by his recent completion of a major project funded by the Department of Electronics & Information Technology, Government of India, focusing on "Developing a New Distributed Computing Solution for Data Mining."

As a testament to his visionary leadership, Prof. Goyal is a founding member of the Advanced Data Analytics and Parallel Technologies (ADAPT) Lab at BITS-Pilani's Pilani Campus, where he currently serves as the head.

In summary, Prof. Navneet Goyal's exceptional contributions to academia, coupled with his pioneering research endeavors and exemplary leadership, underscore his status as a distinguished figure in the field of computer science.



## *Brief Biography of the Co-Supervisor*



**Prof. Poonam Goyal** holds the esteemed position of Professor in the Department of Computer Science & Information Systems at Birla Institute of Technology & Science, Pilani, Pilani Campus. She serves as the coordinator for both the APPCAIR, AI Research Centre, Pilani Campus, and the Web Intelligence and Social Computing Laboratory (WiSOC Lab) within the department. Additionally, she plays a pivotal role as a core member of the Advanced Data Analytics and Parallel Technologies Laboratory (ADAPT Lab).

Prof. Poonam's academic journey encompasses a ME degree in Software Systems from BITS Pilani, followed by a Ph.D. in Mathematics from IIT Roorkee. Her research interests span across diverse domains, including Big Data Analytics, High-Performance Computing, Multimedia Retrieval, Computer Vision, and Natural Language Processing. Through her scholarly endeavors, she has made significant contributions to various social and scientific disciplines, such as Social Media Analytics, Multi-Modal knowledge graphs, Low Resource Languages, and Computer Vision.

Her prolific research output is evidenced by numerous publications in esteemed conferences and journals, including IEEE Transactions on Multimedia, ACM Transactions on Multimedia Computing Communications and Applications, and IEEE Transactions on Social Computing, among others. Prof. Poonam's innovative work has also led to the filing of several patents related to her research across disciplines.

In recognition of her outstanding contributions to the field, Prof. Poonam has received prestigious awards, including the 2021 Google AI for Social Good research award and the 2010 IBM Research Innovation Award under the Smarter Planet Initiative for her work in Scalable Data Analytics. She has also been honored by "India AI" as one of the eight leading women AI researchers in India on International Women's Day 2021.

Furthermore, Prof. Poonam actively participates in various program and review committees for conferences and journals and has served as Principal Investigator (PI) or Co-Principal Investigator (Co-PI) for several sponsored research projects. Her dedication to advancing research in AI and data analytics underscores her status as a prominent figure in the field.





## *Brief Biography of the Co-supervisor from Infosys*

**Dr. Vinay M R** has over 18 years of post-Ph.D. corporate experience, and brings a wealth of expertise in the fields of Data Science, Machine Learning, Predictive Modeling, Advanced Analytics, Big Data along with Economic and Market Research. He has successfully spearheaded numerous large-scale analytical and research projects, delivering strategic business solutions and playing a pivotal role in project planning.



Dr. Vinay's industry and domain expertise span a wide spectrum, encompassing sectors such as automobile, banking & finance, retail, oil & natural gas, telecommunications, entertainment and utilities. Additionally, his functional expertise extends to areas including Campaign Analysis, Churn Analysis, Credit Risk Modeling, Customer Acquisition, Demand & Sales Forecasting, Cross/Up Sell Opportunities, Customer Analytics, Market Expansion, Optimization, Pricing Elasticity & Optimiza-

tion, Revenue Management, and Sales Performance.

Dr. Vinay's academic journey includes a Ph.D. in Economics from the University of Mysore, where his research focused on the impact of macroeconomic reforms on India's manufacturing industry. He also holds a Master of Business Administration (MBA) in Finance from the University of Mysore, where he conducted his research on the impact of financial reforms on the banking sector of India. Furthermore, he also possess a Master's Degree in Economics from the University of Mysore, with a specialization in Econometrics, Quantitative Economics, and Statistics.

With a blend of academic rigor and extensive industry experience, Dr. Vinay has contributed valuable insights and strategic solutions to address complex real word challenges in the fields of data science, analytics, and natural language processing.