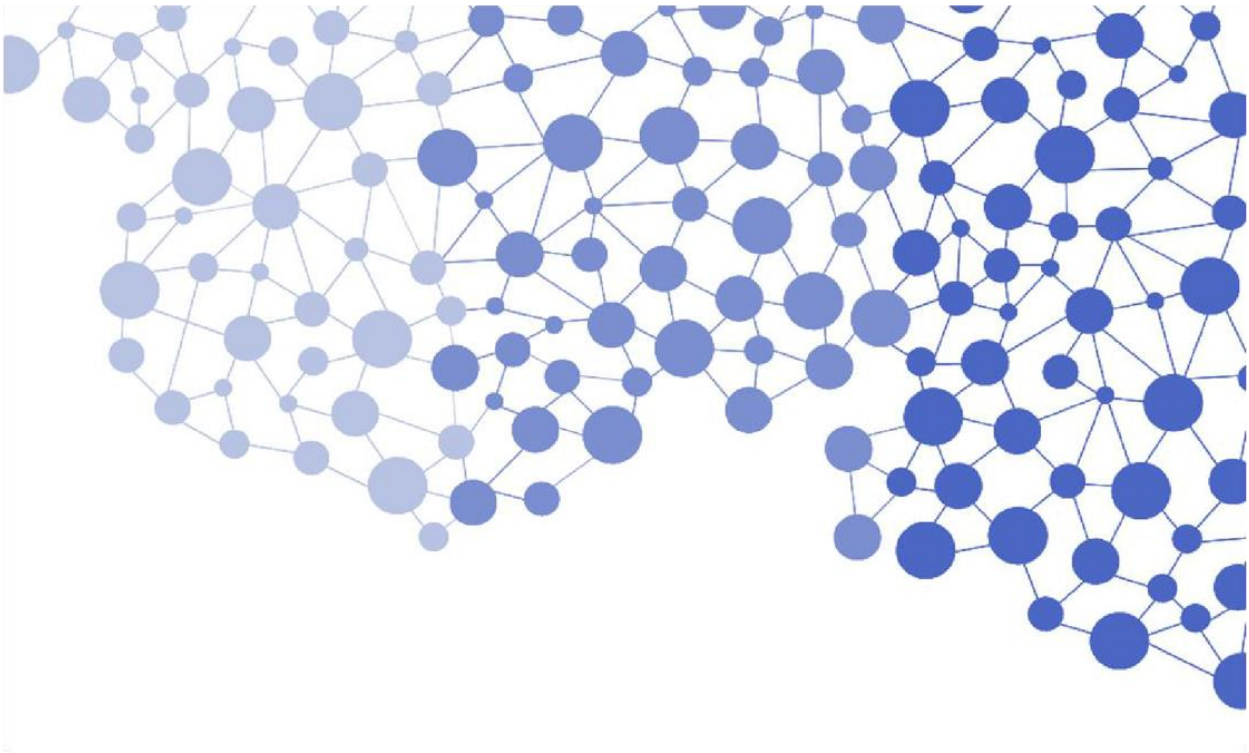
Ministry of Electronics and Information Technology
Government of India



India's Language Technology Transformation

An Evaluation Study of TDIL Programme of Government of India

Dr. Charru Malhotra

DISCLAIMER

This impact assessment report by no means has any commercial intention and is solely undertaken for the purpose of assessing the impact of Technology Development for Indian Languages. IIPA will not be responsible for any interpretation drawn by the reader on the basis of information contained herein. The reader is solely responsible for the use of this material to achieve its intended results. Unauthorized publication/edition/modification of the content of this report is strictly prohibited.

---IIPA Team

ACKNOWLEDGEMENTS

We express our tremendous and sincere gratitude to Chairman-IIPA, Honourable Sh. T. N. Chaturvedi, IAS for staying our never flickering beacon of guidance in this project. Deepest appreciation for our most respected Director Sh. S.N. Tripathi, IAS whose contribution in stimulating suggestions and encouragement helped me to coordinate my project especially in writing this report. Special thanks to Sh. Amitabh Ranjan, Registrar, IIPA, for extending all the institutional and administrative support throughout the project.

We are grateful to Secretary, MeitY, Sh. Ajay Prakash Sawhney, for giving us an opportunity to perform the impact assessment of Technology Development for Indian Languages (TDIL). His persistent support deserves a special mention with all our humble gratitude. We are extremely thankful for being provided such an endless support and guidance, although he has a busy schedule.

We would like to express our deepest appreciation to Additional Secretary, MeitY, Sh. Pankaj Kumar, for his valuable inputs and constant encouragement that helped us in this daunting project. We are grateful for his timely and unconditional guidance till the completion of our report. We are fortunate to have worked with his on this assignment.

We are immensely grateful to the Project Head of MeitY, Dr. S. K. Srivastava. TDIL has had the benefit of the vision of Dr. Srivastava since its inception. He played a major role in bringing the initiative 'National Roll-Out Plan' for wider proliferation of Indian language Software Tools and Fonts. His vision for TDIL helped fuel the growth of industry in this sector. The spin-off of these efforts has resulted into increasing interest of MNCs to look at India as a large market for Language Technologies. India is, thus, poised to emerge as Multilingual Computing hub.

From TDIL team - special mention of Mr. Vijay Kumar for his dedicated support, input and kindest cooperation that made us more surefooted in this convoluted project. Additional thanks to Mr. Bharat Gupta, without his support, this report would not have been so updated and complete.

We can't forget our concealed force of Sh. Mithun Barua, Dy. Registrar, Academic Support, Sh. O.P. Chawla, Dy. Registrar, F& A and Sh. R. D. Kardam, Assistant Registrar, Accounts, Pension & Membership.

As a team, we stay unhesitant to state, "We are indeed incomplete without our leader Dr. Charru Malhotra, who, with her prudent thought process and effective leadership, helped us sail this boat in a distinctive manner."

Special thanks to indispensable team members of IIPA involved in this project. Profuse thanks to Mr. Atul Garg and his team, who burnt the midnight oil and brought shape to this report. He was constantly supported by Aniket Basu, Research Officer, IIPA, for strategic coordination and threading the report in one unit throughout the project along with other IIPA team members. He was wholeheartedly supported by the content team including Ms. Anushka Bhilwar, Mrs. Sugandha Vihan, Mr. Arun Ramasubramanian, Ms. Sanjana Ahluwalia, Ms. Vinti Manchanda, Ms. Nishtha Agarwal, Mr Swentanshu and Ms. Anu Raj Rana. We thank Mr. Amit Garg, Finance and Operations Expert for his continuous support in various tasks of the report generation. Special thanks to Mr. Yatharth Chaudhary for the team building efforts. Last, but not the least, the report

in its beautiful format could not have been possible without the hard work of Mr. Gurpreet Singh and his team.

Printed and Published by

Indian Institute of Public Administration (IIPA)
Indraprastha Estate, Ring Road, New Delhi-110002.
Fax.(O) +91-11-23702440, +91-11-23356528
E-mail: contact_us@iipa.org.in

INDEX

LIST OF FIGURES

LIST OF TABLES

Executive Summary

Before the conceptualization of Technology Development for Indian Language (TDIL) initiative came into place most the of the translation work was done through human labour. From translation to interpretation of the language, everything was dependent on the knowledge of a single person, however, it had its own limitations in terms of its accuracy and applications to ICT. Major ICT applications are pre-dominantly available in English language. However, in multilingual countries like India, where English speaking population is about 5-7 percent, the digital world is still yet to bloom for the non-English speakers. With the initiation of language processing in India, TDIL aims to make "knowledge available for all and digitally unite everyone". Keeping this in mind Government of India, MeitY envisioned of making information technology tools available in 22 officially recognised Indian languages. Various tools were then developed; machine translation, building a corpus, Text-to-speech, speech recognition tools and more to bridge this gap of digital divide based on the language barriers. It has the ability to become the backbone of the Digital India initiative which aims to empower its citizens by improving government services to be made readily available electronically and improving its infrastructure to make the country digitally

empowered in the field of technology. TDIL initiative has the capacity to further facilitate and fuel e-Governance initiatives of the country by engaging more citizens from the grassroot level to be more participatory and aware about government services, education, healthcare facilities, agriculture and banking. With the TDIL project about to complete almost two decades of its inception, Indian Institute of Public Administration (IIPA), New Delhi, was entrusted to undertake impact analysis of the current phase of TDIL.

To do so, IIPA proposed a conceptual framework to analyze the development of individual research areas, tools developed, and its further application based on five essential parameters viz. Technology, Economy, Legal, Operational and Social which were triangulated viz., programme vision with management perspective and research team perspective. To accomplish this, IIPA not just referred the available documents, official reports including project reports, technical evaluation of individual tools but also conducted multiple structured/semi-structured interviews with TDIL team and experts, as well as conducted a detailed primary study with a Questionnaire Survey with the research team of TDIL.

After this comprehensive and methodological evaluation, IIPA concluded that TDIL has become an indispensable part to the various fields of ICT, as well as for the government. Observing the vital contribution of TDIL, IIPA lauds its potential and recommends the continuation of TDIL.

IIPA has also recommended that TDIL in its next phase of initiation should strive towards making individual tools more impactful at the national level than what it is today. It should continue to extend its research and development in order to make individual tools more accurate and accessible. Furthermore, its primary objective should be in regard to enhance the process of commercialisation and deployment, wherein, it encompasses more and more institutes, organisations, start-ups and large company players to work in collaboration with the TDIL project. The corpus should be readily made available in public domain for various start-ups and independent researchers at a lower cost or free of cost. Furthermore, by effectively setting up a large and rich start-up ecosystem TDIL can provide more managed services to its citizens and facilitate India's economy in the global context.

The table below shows the recommendation from the perspective of each pillar based on the approach used for the study.

Table 0. 1: Recommendations on the TELOS framework

FRAMEWORK PILLAR	RECOMMENDATION
Technical	1. Rethinking and Redesigning the Corpus a. Current Corpus size of Indian languages is very small for the effective development of machine translation and language technology developments. b. The idea is to develop our Corpus, which could be understood as the building block for this whole TDIL programme. c. Government already has a data sharing policy (2014) in place and a chief data officers is already designated in each organisation, ministry and department. These officers shall be given the responsibility of providing the data of that department as per the data sharing policy 2014 for the

	language development programme. d. Various Government departments have huge amount of current and archive data i.e., Doordarshan, National Archives of India, different ministries etc., can be asked to provide the data.
	2. Deployment and New Architecture of DC platform a. In order to increase the usage, the developers of TDIL should more focus on the deployment of tools and services of Indian Languages. b. We can plan further to make the next versions of DC platforms to be built like that of new age configuration cloud infrastructure platforms where Indian Language infrastructure, services and components are readily available on cloud infrastructure. c. From the developer's point of view, we can focus on the components of language technology in a "plug-n-play" kind of architecture where applications can be built by dragging and dropping the components.
Economic	1. Market Creation for Language Technology and Application by Central and State Governments and Industry a. Central and State Governments should help in creating jobs and business opportunities in the applications of Indian language technology. b. Existing large companies and commercial players could also be convinced to provide products, services and platforms in Indian Languages catering to non-English speaking population of India.
	2. Building a Start-up Ecosystem a. In the field of language processing and technology start-ups can be quite beneficial in terms of commercialisation process, getting more manpower and innovative ideas to digitally empower the citizens. b. Start-ups in Indian Languages providing different technology services and applications need to be promoted by the ministry. Seed funding should be provided along with the incubation support, marketing exposure and the technology tools to flourish the business environment.
	3. Efficient and Continuous Funding of the Programme a. To further streamline the processes of developing our research and development methods and tools there is a need of efficient funding for the project. The funds thus also need to be released timely by the respective authorities.
Legal	1. Open Source and Free Licensing a. The technologies and components developed under TDIL should be open source and available free of cost to users and development community. The technologies should be available under GNU General Public License (GPL) for further use and development. The open source community can develop the technology and build multiple solutions for the industry and users.

Operational/ Organisational	**1. Dictating the Vision and Mission** a. In order to initiate a programme, there is need to define set goals and aspirations for the project under a well-defined, coherent, hierarchical structure which by its virtue should empower individual projects through adequate focus and allocation of resources. b. A roadmap needs to be created with a set number of tasks in hand pertaining to each phase of development. c. Creation of timeline should be done to one, overcome the current problems that have arisen by taking up large number of tasks at one-go. And two, will help in defining the goals for better efficient outcome.
	2. Content Creation a. Central and state governments, organisations and departments should aim to create independent content in Indian languages for various domains and make it available online on different platforms. b. The central and state governments need to outsource the creation of content to begin with, to expand the information available online in Indian languages.
	3. Project Management Unit a. Call for a continuous monitoring with alerts and notifications so that management has granular visibility of task and can see the green, yellow and red flags against the project. b. Effective management can also be implemented through operational dashboards linked with accountability metrics and a risk management plan. An accountability plan can be an important feature for the entire life cycle of the TDIL programme as it could be used at different stages to keep a check on its progress.
	4. Government and e-Governance Sites a. Language processing and translation of Indian Languages should be first implemented on government sites and e-Governance services and mobile apps. Secondly, in light of the same, Data Centre Portal could be made available in all TDIL supported languages, at least to start with.
	5. State Language Missions a. To build the TDIL corpus more efficient several state IT secretaries should be invited on board. They should be sensitised and pave the way forward into leading to the creation of a Unit in their respective state's IT department or the selected state language body. b. Specialised training and awareness campaigns should be conducted for officials along with focusing on deployment of language technologies in state e-government services in collaboration with participatory start-ups. c. Sub language projects/missions for various individual languages could be set in place.

	6. Continuous Evaluation and Assessment with International Standards a. With far wider scope of work and improvement, the government thus has a far better and a fortunate opportunity to learn from its International peers and collaborators. Evaluation should be a continuous process along with benchmarking with International standards.
	7. Awareness Branding and Public Relations Unit a. TDIL not only has a greater impact on the digital advancement of India but a far greater impact on its social and economic growth. The need is to realise and make people more aware about these initiatives which will allow them to be knowledgeable and aware about such facilities.
	8. Support System with Escalation Mechanism a. An escalation plan needs to effectively set in place to deal with potential problems such as for developing individual tools, assisting the public with the working and functions of the tools, commercialisation support, start-up support etc.
Social	1. Development of Human Resource a. Inclusion of Language Technology in academics can clear the path of the early adoption of the Language Translation. b. Training programmes and workshops are required at frequent intervals for the officers and the torch bearers in different geographies. These programmes should be extended not only for the technology and the tools usage but also for the next level of technology, services and platforms development along with the commercialisation and go to market training. c. Online programmes can be created by the TDIL in partnership with some of the training development organisations i.e. C-DAC and related training institutions and made available to the all the stakeholders.
	2. Sector Wise Products and Services a. We can further enhance the information and service environment by making the developing tools available for different sectors.

Way Forward

The TDIL has thus since its inception set up its foundation in the field of Indian Language technology by striving to make its more and more accessible to the masses and through its robustness, scalability and inventive nature established its potential to contribute to the country's economic growth. The IIPA in this regard has recommended a 'Commercial Business Unit' for the TDIL to help the programme reach its true potential. This Business unit should be an interconnected one which links the needs and aspirations of the Citizens along with the R&D in the area and the various capacity building innovation platforms and collaborations. This will aid the creation of an ecosystem which will allow for further development and reach for the programme. It will also facilitate the ability of this business unit which has the complete

information about the various needs and existing applications for Indian Language technology, to look after the complete building capacity for technology under one umbrella from education to training. This change in methodology the IIPA suggests will also promote the development of tools for the growth of various sectors of Indian economy such as banking, tourism, agriculture, education etc. There is also the need to take into account the newer technologies which will help build fresh products, services and platforms. A shift from the technology infrastructure model is also proposed to bring about a more holistic approach to the programme, which would help provide an empowering platform for enhancing the knowledge economy of the nation. After this comprehensive and methodological evaluation, IIPA concludes that TDIL can become an indispensable part to the government and R&D and sectoral community for empowering the citizens. Observing the vital need of TDIL, IIPA lauds its potential and recommends enhancing its research, deployment and commercialisation processes along with the right funding by taking in account the future of this technology and empowerment that it can bring to the citizen.

Chapter 1: Introduction to the Study

1.1 CHAPTER OVERVIEW

This chapter delineates the objectives of the third-party evaluation of TDIL. Starting with a brief introduction to study the chapter then moving on to defining the need and the scope of the study. The subsequent sections talk about the five stages, viz., Inception Report, Detailed Action Plan, Submission & Development of the Assessment Framework, Draft Report and Final Report on Impact Assessment. The chapter further defines the methodology of the study under the headings of "Defining the Assessment Framework" and "Stages of the Evaluation Study" as well as the governing principles of the assessment that serves as a guiding template to the study.

1.2 INTRODUCTION TO THE STUDY

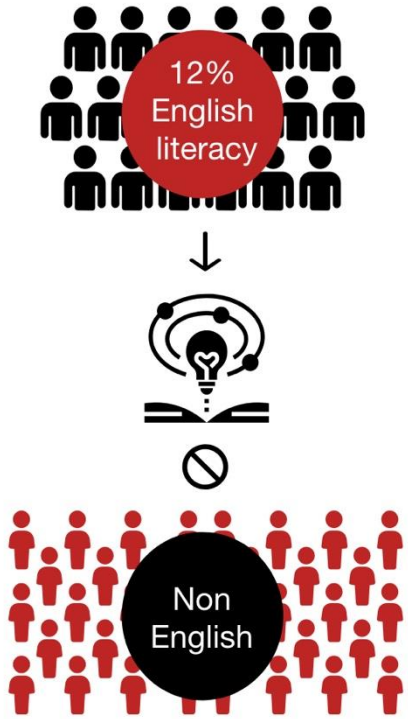


Figure 1. 1: English literacy divide

Study of the Impact of the TDIL (Technology Development for Indian Languages) programme is a challenge as the programme has a history of more than 33 years and has undergone multiple transformations (It was started in the year 1986 under MeitY and became TDIL in the year 1991). This assessment is imperative as it showcases the importance of this programme in enabling and empowering the citizens of India to access and utilise the information beyond those available in mainstream global languages like English and Hindi. India is unique in its linguistic heritage with 4 language families, namely, Indo Aryan (76.87 % speakers), Dravidian (20.82 % speakers), Austro-Asiatic (1.11 %), and Tibeto-Burman (1%).Beyond this, a new language family called 'Andamanese' has been recently discovered (Abbi 2001), and there is a possibility of another – a 6th family - called 'Great Andamanese'. There are 22 constitutionally recognised languages and more than a thousand dialects in the country. Hindi is recognised with having both the 'official' and the 'national' status while English has the status of 'associate official' language.

Figure 1. 2: Recognised languages in India

The need of TDIL is rooted in the fact that most of the textual information available to the citizen through ICT is either in English or in Hindi. This is a hindrance to the development of the country as well as the common citizen since they are unable to comprehend the information available to them, due to it not being available in their mother tongue. This has a lasting impact on the skill set and employability of the youth in the country. Only 12% of the Indian population is literate in English (fig 1.1), yet most of the information available on government portals, educational websites and the internet, in general, is in English. This language barrier becomes a major hindrance not only for governance purposes but also for education and the overall development of the population. TDIL programme has contributed towards decreasing the gap in terms of accessibility to information for people of the varied vernacular in the country. Exchange of information at a global level is a key to holistic development of all levels in society. Hence, an extensive study of the role of TDIL in some key sectors, viz., education, healthcare, agriculture, technology, science and governance, is a matter of importance.

1.3 SCOPE AND OBJECTIVE OF THE STUDY

TDIL was initiated in the year 1991 under the Ministry of Electronics and Information Technology (MeitY) and had the objective of developing Information Processing Tools and Techniques to facilitate human-machine interaction without a language barrier, as well as creating and accessing multilingual knowledge resources and integrating them to develop innovative user products and services.

The scope (fig 1.3) includes the assessment of the TDIL programme regarding the objectives as envisioned from time to time. There is also need to compare the technologies that have been developed under the programme with some of the technologies available for other languages in the world, as well as the extent to which the technologies that have been developed and utilised by some of the identified user-agencies or commercialised by industry.

Figure 1. 3: Scope of Impact Assessment

The assessment tries to discover the constraints in the utilisation or commercialisation of the developed technologies and suggests the steps to mitigate the same. The scope includes undertaking stakeholder consultations with the relevant agencies and conducting site visits wherever felt necessary.

The scope also analyses how the programme has helped in the proliferation of the content and applications in Indian languages and to see how these have benefitted the intended beneficiaries. There is also a need for scrutiny of the work done for the development of standards in Indian languages and suggest further steps for improving the same. The need to identify gaps in the achievement of the programme and suggesting steps, modifications and improvements for the same also encompass the scope of the study. Finally, suggesting directions for initiating new activities to utilise the opportunities being offered by the recent advances in technology.

1.4 DEFINING THE ASSESSMENT FRAMEWORK AND METHODOLOGY

The impact assessment revolves around the general principles of reviewing the past work that has been done, using an assessment tool to gauge the Technical, Economic, Legal, Organisational/Operational and Social impact of identified research areas that encompass the multitude of projects done under the TDIL programme. There will also be a questionnaire-based survey tool that will be used to gauge the key positives, recommendations and concerns regarding the programme. Lastly, using these assessment tools, key recommendations and the way forward have also been proposed.

To achieve the aforementioned objectives, IIPA has customised an assessment framework for TDIL evaluation. The framework contains the following architecture - the first step was to understand the TDIL vision and its objectives which has been dynamic in style and has been evolving because of the R & D nature of the programme. Next was to understand the TDIL management perspective and research teams' perspective. The objective in the study was to get the validation of the management perspective by the research teams and vice versa. Both perspectives in turn validates the TDIL vision. The framework used in the process was TELOS (Technical, Economic, Legal, Operational and Social) based analysis, primary research of the system and getting the comparison by validating it against the global best practices. The methodology used is depicted in Fig 1.4.

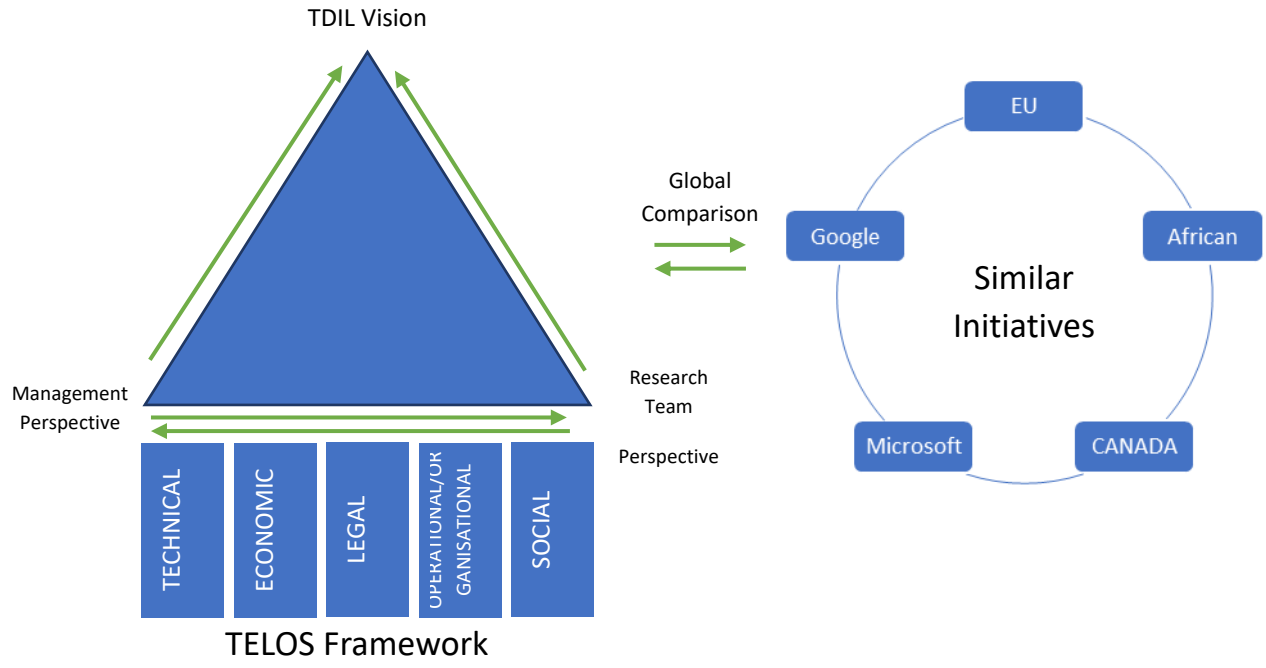


Figure 1. 4: Methodology of study

The foremost objective of the TDIL evaluation study was to study the achievements of the programme with regard to its vision from time to time was captured and validated from a management as well as institutional perspective. The whole framework of the assessment

study was designed to achieve the same. The evaluation also has an entire chapter dedicated to to the study and comparison of the different language technologies around the world to that of the Indian language technology programme. The chapter includes the study of private language technology initiatives such as Microsoft and Google as well. One complete chapter has been dedicated to the study of the extent of the utilisation and commercialisation of Indian Language technology. It encompasses the study of various applications and sectoral initiatives under the programme. The constraints of the same have also been touched upon and included in the Chapters 3 which deals with the Fundamental Research areas under the programme along with Chapter 6 which deals with the Commercialisation Efforts of the same. The chapter on Recommendations suggests steps to overcome and mitigate these constraints.

After studying the system from the perspective of a casual user we conducted a survey for the primary research analysis of the TDIL programme with various institutes and Technical Research teams involved. There were also several face to face meetings with the various researchers of the same to facilitate a better understanding of the programme. The team additionally conducted site visits to the Ministry as well as the C-DAC office at Noida to assess the ground level situation for TDIL. The analysis of the Proliferation of the various contents and applications of Indian Language technology has been studied in Chapter 5 of the evaluation that deals with the deployment of the TDIL portal. The beneficiary impact of the same has been discussed in the chapters dealing with Findings, Survey and Recommendations. The work done on development of standards for Indian Languages has been taken up in Chapter 4 of the assessment study which examines the journey of Indian Language standards as well as the various International standards that are currently being used. The chapter on Findings deals with the different steps that can be taken to improve the same.

The gaps in the TDIL programme along with the suggested modifications and improvements for the further promotion of the same have been reviewed in the chapters dealing with Findings, Survey and Recommendations. Finally, the directions for initiating new activities in order to make the best use of the opportunities that recent advancements in technology offer for the TDIL programme have been covered under the 'Technology aspect' part in the chapter on Recommendations.

1.5 STAGES OF TDIL IMPACT STUDY

The course of the study is segregated into five stages (fig 1.5). Each stage is specifically identified through its deliverable, viz., Inception Report in Stage-I, Detailed Action Plan in Stage-II, Submission of Assessment Framework, Development of Data Matrix and Assessment Framework in Stage-III, Draft Report in Stage-IV and Final Report on Impact Assessment of TDIL in the last stage.

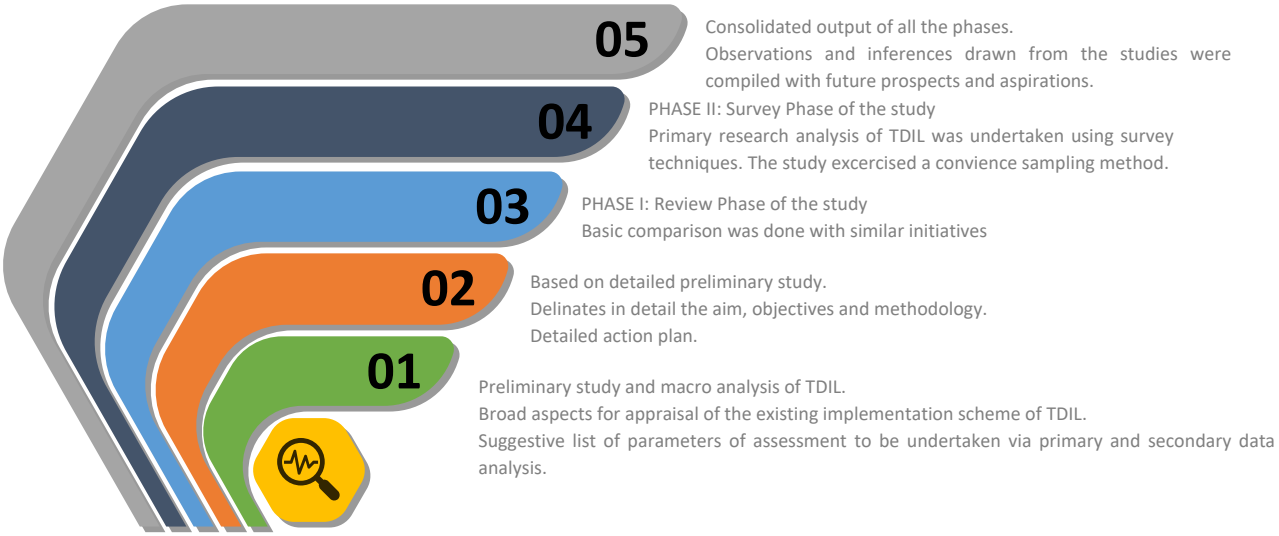


Figure 1. 5: Stages of impact assessment

Stage I: The deliverable of the first stage, i.e., Inception Report, was developed after the preliminary study and a macro level analysis of TDIL. This report articulates the broad aspects for appraisal of the existing implementation schema of TDIL and presented a suggestive list of parameters for the assessment to be undertaken through primary and secondary data analysis.

Stage II: The second stage of the study, i.e., Detailed Action Plan, was based on detailed preliminary study, undertaken to delineate in detail, the aim, objectives and methodology that is to be followed for undertaking the study along with the detailed action plan.

Stage III: Further, to move on to Stage-III, the study was conducted in two phases, viz., the review (Phase-I) and detailed research analysis (Phase-II).

Phase-I: The Review Phase of the Study - The first phase included an overview of the existing technologies and services as an outsider, then from the macro-perspective of institutes and other stakeholders, including management and citizens. In this phase, a basic comparison was done with similar initiatives prevalent elsewhere globally.

Phase-II: The Survey Phase of the Study - After understanding the system from the perspective of a casual user, it was deemed important to take the rightful beneficiary agencies' opinion on the same. For doing so, a primary research analysis of TDIL was undertaken using survey techniques.

Sample Used: The study exercised a convenience sampling method to obtain responses from institute developers of TDIL and its programmes.

Stage IV: The Stage IV is the consolidated output of all the above-mentioned phases. Here, all the data is drawn from the review of systems and primary research that was collated. The observations and inferences drawn from the studies were complied with future prospects and aspirations.

Stage V: The study was conducted with a vision to formulate the final report on impact assessment of TDIL. The report is extensive in its reach and intensive in its context, deriving the strengths and crucial insights into this vast knowledge network and a service delivery platform.

1.6 DELIVERABLES AND STRUCTURE OF THE STUDY REPORT

This report is an outcome and an overview of the results attained through the review and primary and secondary research analysis undertaken in the study, the deliverables of the study are depicted in Fig 1.6. A birds' eye-view of the same is represented herein:

Chapter One - "Introduction to the Study: This presents the need and background of the study and further it gives scope of work along with an overview of the methodology and framework used to pursue the proposed impact assessment study.

Chapter Two - 'TDIL: Need, Vision & History': This chapter covers the need, vision, genesis, and architecture of the TDIL Programme.

Chapter Three - 'Understanding TDIL: Fundamental Research in language Technology', This lays out the multi-layered structure of the TDIL programme, that helps us visualise the different aspects of the programme enabling us to understand the fundamental research areas and development of same over a period of time.

Chapter Four - 'Understanding TDIL: Standardisation in Language Technology' covers the collaboration of TDIL with different National and International organisations to develop Standards for Indian Language Technology. This builds up the baseline and solid foundation for the language technology research in India.

Chapter Five – 'Understanding TDIL: DC Portal and Deployment' chapter entails the portal architecture, available applications and the security aspects. Detailed dashboards of resource usage and portal visits are also described here.

Chapter Six – 'Understanding TDIL: DC Portal and Deployment' chapter entails the portal architecture, available applications and the security aspects. Detailed dashboards of resource usage and portal visits are also described here.

Chapter Seven–'Review using the TELOS framework': This chapter covers the analysis of the programme using a defined structure, TELOS - Technical, Economic, Legal, Operational/ organisational, and Social. Using this structure, the chapter reviews the TDIL programme and its projects, in order to define and assess their status and impact.

Figure 1. 6: Project deliverables

Chapter Eight–'Survey analysis and Feedback': This chapter aims to convert the output of the primary research analysis of TDIL gleaned through survey technique employed in the study.

Chapter Nine–'Global Benchmarking with Similar Initiatives': This chapter assesses the Global Language Technology initiatives which are similar in nature to TDIL.

Chapter Ten–'Key Observations & Recommendations': This chapter integrates all the findings collected through review and primary research conducted in this study. Based on these observations and analysis of consultation rounds, several relevant recommendations are put forth for strengthening important facets of TDIL including its operational, technical, organisational functions and its sustenance.

Chapter 2: Background of TDIL

2.1 INTRODUCTION

The Ministry of Electronics and Information Technology (MeitY), Government of India, recognising the need for good governance that would include effective implementation and interaction between the government, stakeholders and the citizens, had decided to launch the initiative of Technology Development for Indian Languages (TDIL). This programme aims to bring various information processing tools and modern technologies into everyday practice for various Indian Languages, so that they can be used easily by a common person in order to derive benefits from ICT.

India is currently in the phase of expanding its reach in the field of information technology by making its various applications user friendly. Being a multilingual country, it was realised that for these various IT applications to be citizen centric, there was a need for them to be readily available in one's own language. Most of the advanced technologies were originally developed using English and were subsequently established in other language interfaces due to the multi linguistic nature of the world. When it comes to languages, people are more comfortable in their mother tongue as a form of communication between people or working with modern technologies. This essential need thus made the development of IT applications into various language interfaces imminent. Adhering to this, initial attempts to develop a machine translation system were made in the early fifties. India, following these efforts set up its own area of expertise that worked on language technology initially focusing on machine translation. These attempts were undertaken under the banner of Knowledge Based Computer Systems (KBCS) programme in 1986 and later under the Technology Development for Indian Languages (TDIL) in

1991. These programmes were advanced with the aim of developing tools for information processing in order to facilitate human machine interaction using Indian languages and building up technologies to access multilingual knowledge resources.

2.2 HISTORY

In 1949, the Rockefeller Foundation presented a computer-based machine translation that was comprised of information theory, code breaking during the Second World War and theories about the universal principles underlying natural language. The system was referred to as a "toy" system comprising of only 250 words from the field of Chemistry and translated into 49 sentences from Russian to English.

The model, however, was good enough to give an idea and a way forward for future applications and developments. Since then, different countries have been putting efforts to develop machine translation systems for different interests. While acknowledging the various complications it was realised that machine translation system can be quite helpful in improving the productivity of human translators. With the growing advancements in the field of IT and by allowing it to be more than just technology for professionals, a feature for all, the need of translation grew more and more. This is also the reason why countries across the world continue to support R&D in this particular field and facilitate efforts to make fully automatic machine translation a reality.

India too has been working in the field of language technology including machine translation keeping in view the language diversity of our country. The main objective was to make services readily available in various local languages, starting with Hindi. The Ministry of Communications and Information Technology (MCIT) started working on language technology in 1986 with a special focus on machine translation, keeping in mind the multilingual nature of our country. This programme, known as the Knowledge Based Computer Systems (KBCS), set up a centre for research in speech technology at the Tata Institute of Fundamental Research, Bombay, and C-DAC, Pune, which was formerly known as the National Centre for Software Technology that started working on the Natural Language Processing especially in regards to machine translation. Subsequently, the Ministry in 1991 initiated Technology Development for Indian Languages to build more on the research and progress in this area to come up with technological solutions for Indian Languages. This programme intended to develop information processing tools and techniques to create knowledge resources that could be made available in various languages for the enhancement of human interaction and for the development of user centric products.

2.3 NEED OF TDIL

Figure 2. 1: Languages as a barrier to accesses ICT tools and services

Majority of the people in India are bilingual, wherein English is stated to be the second language. The English literacy level in India, however, is quite low as a result of which a large percentage of the population is not able to tap into the benefits of the internet. Language, then, is a barrier that halts the democratisation of data and hampers the development of an efficient system of public governance (Fig 2.1). In this day and age of internet revolution, where a majority of the work done is through online means, barring the use of WhatsApp, there is hardly any smartphone usage which could be recorded. The primary reason for such a limited usage is due to the language barrier, leading the TDIL programme towards its major objective to make this interface available to the people in their native language.

Figure 2. 2: Components of Good Governance

Not using one's native language for communication can be observed as a deal breaker in context of various economic, social and political progress initiatives. If the already available text information and access to the internet and other e-governance programmes were also available in one's known language, the benefits of the same could be utilised by most of the population enhancing the progress of the citizens and the country as a whole. Furthermore, language is the key to communication which is important in every strata of the society. In order to build a progressive and developing society, communication is a must for every established or establishing governance. The process of governance is made up of the interactive communication between the citizens, the stakeholders and the governing body. Thus, communication is imperative and can be regarded as an important element of good governance. A further detailing of what good governance entails can be found in Fig 2.2. In a

democratic country like India, the most utopian form of good governance relies on a responsive framework. In order to achieve that, it requires not only a stronghold on IT but also an effective communication, i.e., eradicating challenges presented by the language barrier. India is a multilingual country, with 22 constitutionally recognised languages and 11 scripts. While most of the work, especially in the field of IT, is done in English, according to the stats only 5-7 percent of the population of India can speak and write English. As a result, a larger part of our society is left deprived of the developments in this sphere.

In today's time where IT is breaking the barrier and bringing the government and the citizen closer, it becomes essential to enable a wide proliferation of Information and Communication Technologies in various Indian Languages.

Under the Digital India initiative, the primary aim is to provide multilingual ICT based solution development in Indian Languages. In order to make the most of IT, it is necessary for its services to be readily available to the masses. Even in situations where the services are available, it is required for them to be easy to understand for the citizens. In this context, language, thus, plays a key role to curb the problem of accessibility of these services to the people. Moreover, multilingual ICT based solution is a leading component in facilitating Digital India Initiative, namely:

- e-Kranti: It emerged in the background of the critical need to transform e-governance in order to promote mobile governance and good governance in the country. It envisions to transform e-governance to facilitate a transformation in governance.
- Information for all: To make information available for the citizen to enhance its accessibility and improve transparency between the government and the citizen. Datasets are released by the ministries/departments to be used, reused and redistributed.

Furthermore, TDIL programme promotes standardisation of Language Technology in the areas for seamless access of information across platforms. Thus, development of machine translation shall not only benefit the citizens in terms of good governance but also in the development of various domains such as education. To improve academic performances of the students, the National Programme on Technology Enabled Learning (NPTEL) was launched with the aim to translate the academic texts available in English to Indian languages. A large section of students who come from rural background tend to drop out of professional courses because of the language barrier as the medium of instruction is mainly in English. There is also a lack of availability of study material in local languages. This not only affects their education it also serves as a tool for denying them their fundamental right to education.

Figure 2. 3: Language Technology bridging the access gap to ICT benefits

Besides education, machine translation can also be quite effective in the field of healthcare, agriculture and law. Language technology, if made easily available, can help the masses transcend the aforementioned language barrier and access the various benefits of ICT services (Fig 2.3). Availability of information in one's local language serves to boost up the e-health initiatives by facilitating the accessibility of information. Farmers too can be more informed about the government's key initiatives in the field of agriculture and ease the understanding of these policies as well as other agricultural information. Judicial proceedings could also be sped up, as once the case is transferred from lower courts to higher courts including the verdicts and documents are translated. Machine translation will thus help in ease of translation.

2.4 VISION OF TDIL

Technology Development for Indian Languages (TDIL) Programme initiated by the Ministry of Electronics and Information Technology (MeitY), Government of India, has the objective to develop information processing tools to facilitate human machine interaction in Indian languages and to develop technologies to access multilingual knowledge resources.

Under the flagship of Digital India, the vision is to create information for all without any language barriers. Understanding the need of the citizens the government has come up with an initiative which allows accessibility of information and data beyond language. With this we are thriving to make India a digital country by providing internet services to various parts of the nation. There is also a need to make this information available to the people in their local language and also making this information accessible to the masses in general.

The objective of the programme is, thus, to build high-quality Speech to Speech Machine Translation (SSMT) System for Indian local languages in order to promote the research and development of technology, software tools and applications to facilitate translation. The TDIL also aims to create text and speech corpus in Indian languages which is essential to develop language technology solutions. Establishment of Centres of Excellence will take place at the participating institutions to do the required engineering in order to improve accuracy and efficiency of the existing language technology tools and develop new tools:

1. Development of standards for linguistic resources, tools and applications to create and nurture an ecosystem enhancing interoperability. It includes start-ups, central and state

government agencies. The aim is to build an ecosystem of collaboration to develop and deploy innovative products and services in Indian languages.

2. New initiatives are being taken up with a focus on major areas like the Machine Translation System which translates available data in English to Indian languages or data from Indian language to another.

3. Other methods include development of Optical Character Recognition (OCR) system to convert printable text in editable format. On-line Handwriting Recognition system converting handwritten text to editable text, Cross-lingual Information Access System and Speech Processing System in which local language text can be read by machines and spoken words can be recognised.

The tools and resources which will be developed would be offered at an affordable price to the developers to allow them to build inexpensive solutions in Indian languages, wherein Centre of Excellences (CoEs) will provide the required mentoring and support to the start-ups sunder the mission.

2.5 SALIENT FEATURES OF TDIL

The main objective of the programme is to facilitate localisation for the access of ICT services and products in local languages. The focus is on the development of language computing technologies for all 22 constitutionally recognised Indian languages. The outcomes from the TDIL programme thus works to enhance ICT accessibility.

1. As an initiative, the programme has been observed as a unique initiative by the Government of India and has been recognised globally for spawning research across 22 Indian Languages.

2. It has enabled various nurturing research groups across the country in the development of language technologies.

3. Standardisation of Language Technology has further enhanced the policy measures for interoperable solutions across various devices, platforms and contents.

4. Various international standardisation bodies like Unicode, W3C, ELRA, and GALA, have made notable contribution towards the requirement to incorporate Indian Languages in global standards.

2.6 KEY INITIATIVES

The outcomes of the projects undertaken under TDIL Programme are showcased through the Indian Language Technology Proliferation & Deployment Centre portal. The portal acts as a national repository for linguistic resources, standards, contents of language, tools and applications.

2.7 TECHNOLOGY OFFERINGS

1. Text-To-Speech (TTS) in Indian Languages: It is available as a browser plug-in for Mozilla and Chrome Browser for eight Indian Languages, including Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Odia and Gujarati. The languages have been developed and made available through TDIL data centre.

2. SMS Reader in Indian Languages: Developed for Android Platform and caters to 5 Indian Languages, namely, Hindi, Marathi, Tamil, Telugu and Gujarati. It has been developed and made available through e-Gov Mobile Seva Gateway.

Figure 2. 4: Technology offerings of TDIL

3. Automatic Speech Recognition (ASR): Developed for Agricultural Commodity prices for 6 Indian Languages, namely, Hindi, Bengali, Assamese, Tamil, Telugu and Marathi. The system acts as a voice interface for NIC Agmarknet portal in 11 Indian Languages/Dialects for Agricultural Commodity prices and weather information system.

4. Machine Translation Systems: The systems involve translating English to Indian Languages and Indian Languages to Indian Languages, like English to Bengali, Bodo, Gujarati, Hindi, Marathi, Odia etc., and Hindi-Bengali, Hindi-Telugu, Urdu-Hindi, etc. The service for translation of simple English sentences can also be accessed through mobile SMS and mobile internet.

5. Optical Character Recognition in Indian Languages: To make documents in editable format especially for Indian Languages, Optical Character Recognition for 9 Indian languages have been developed. The system includes languages such as Bengali, Devanagari, Gurumukhi, Kannada, Malayalam, Telugu, Tamil, Assamese and Urdu.

6. On-line Handwriting recognition system (OHWR): Ability of a computer to receive and interpret handwritten sources such as paper documents and photographs have been developed in 8 Indian languages. Hindi, Bengali, Tamil, Telugu, Kannada, Malayalam, Assamese and Punjab. Within this system the Tamil OHWR system has also been developed for Android and win 8.1 and is hosted on TDIL-DC portal.

7. Cross Lingual Information Access (CLIA): Cross-language information access enables users to retrieve information in languages one is fluent in and utilises the basics of language translation methods to retrieve documents originally written in other languages. Monolingual Search Engines for Tourism Domain for 5 Indian Languages - Hindi, Bengali, Marathi, Tamil and Telugu - have also been released on the public domain.

8. Sanskrit Tools: Tools which help to read, transcribe and process Sanskrit texts like the Sanskrit Morphological Generator, Morphological Analyser, Sandhi, Sadhi-Splitter and Transliteration are also hosted.

9. Glossary Tool: A system to especially make a glossary. It includes a list of terms in all the 22 Indian languages.

10. Localisation Project Management System (LPMS): To publish documents for localisation and managing localisation projects a web-based platform has been set up, along with translators and reviewer's dashboard.

11. Language CDs: Various tools of free language CDs are now available in all 22 constitutionally recognised Indian languages. The aim is to enhance Office productivity, application and e-development, and are being used by different PSUs, Banks, Educational Institutions, etc. The CDs are compatible with Windows 8.1 and Ubuntu 11.04. One example of this is the SakalBharati open type font which was released in public domain to enhance creative expression of citizens on the web with ease. Along with it, 22 language keyboard drivers for Android was implemented to help connect the smart phones users with digital world in their own language.

Fig 2.4 offers a quick glimpse into the above technologies.

2.8 RESOURCE OFFERINGS

Annotated Text Corpora: Annotated Text Corpora consists of a wide collection of 100,000 sentences for each of 10 languages comprising of Hindi, English, Bengali, Urdu, Gujarati, Konkani, Malayalam, Marathi, Tamil, Punjabi and Telugu. It is majorly available for tourism and health domain.

WordNet: In order to create niche technologies such as Machine Translation systems, etc, WordNet data is quite useful. It is used for developing bilingual dictionaries and Word Sense Disambiguation module. Currently, it has been developed and made available for eighteen Indian languages such as Assamese, Bengali, Bodo, Gujarati, Kannada, Kashmiri, Tamil, etc. on the TDIL Data centre under the Integrated Indo Wordnet.

2.9 OTHER OFFERINGS

Standardisation: Understanding the need to break the language barrier an apex system for the maintenance of standardisation has been developed. The aim is to achieve communication by standardising Indian languages in the field of ICT.

Validators/Localisation Tools: To achieve a seamless access across platforms and devices, a validation of the webpage source is essential. So far, W3C has developed open source validators to assess validity of Web documents in HTML, XHTML, SMIL, MathML, etc. Some of the validators that are in use are P3P, Validator, XML Schema Validator, Glossary Tool and localisation plugin, etc.

Technical Journal of Indian Language Technologies: To consolidate information about products, tools, services, activities, developments, and achievements of Indian language software, the initiative of creating a technical journal has been taken. The VishwaBharat@tdil is a technical journal which keeps in store the information related to the software in one place and can be accessed through the TDIL website.

Chapter 3: Fundamental Research in TDIL

3.1 OVERVIEW

This chapter aims at developing an understanding about the components of research areas in TDIL Programme by studying the existing system. The understanding of the research areas was gauged through various structured and unstructured documents provided on these components.

The chapter gives a brief introduction of the research areas in TDIL Programme, its seven pillars on which the research of TDIL Programme is focussed upon, namely, development of Optical Character Recognition (OCR), Automated Speech Recognition (ASR), Online Hand-writing Recognition (OHWR), Machine Translation (MT), Development of Cross-lingual Information Access (CLIA) Systems, Text To Speech (TTS) and Text Corpora in machine readable form. The study further explains the various aspects of TDIL covered under each of these seven research areas and shows how each of these have created an impact to strengthen TDIL Programme.

3.2 INTRODUCTION

TDIL programme visions to empower citizens from all classes and background taking the benefits of ICT and knowledge sharing. When we talk about empowering everyone, the various divisions of the society aren't only based on the region and class, but also on the individual abilities. With the aim of delivering services to every individual despite of their abilities and disabilities, were the language technology software developed. For example, Text-to-Speech software delivers machine readable texts and figures into human voice. Thus, once the software is integrated with the screen reader, it shall allow people with visual impairments and reading disabilities to listen to written works on a computer, mobile device or a tablet.

Figure 3. 1: Research areas under TDIL

To achieve its vision, the TDIL Programme is developing various components as seen in fig 3.1.

3.3 RESEARCH AREAS UNDER TDIL PROGRAMME

TDIL is doing fundamental research in the following areas. These research areas are discussed in detail with a view of providing their needs and benefits as well as their usage in TDIL Programme.

3.3.1 DEVELOPMENT OF ROBUST DOCUMENT ANALYSIS & RECOGNITION SYSTEM FOR INDIAN LANGUAGES (OPTICAL CHARACTER RECOGNITION)

Overview

Optical Character Recognition (OCR) is a process that can convert text, present in digital image to editable text. It allows a machine to recognise characters through optical mechanisms. For over a half century, research in this area has been ongoing and character recognition rate in modern OCR is producing very good results both in printed text recognition on a high-quality document, and on handwritten documents. First patents on OCR were registered in 1929 and 1933 while the first commercial computer for OCR came out in 1951.

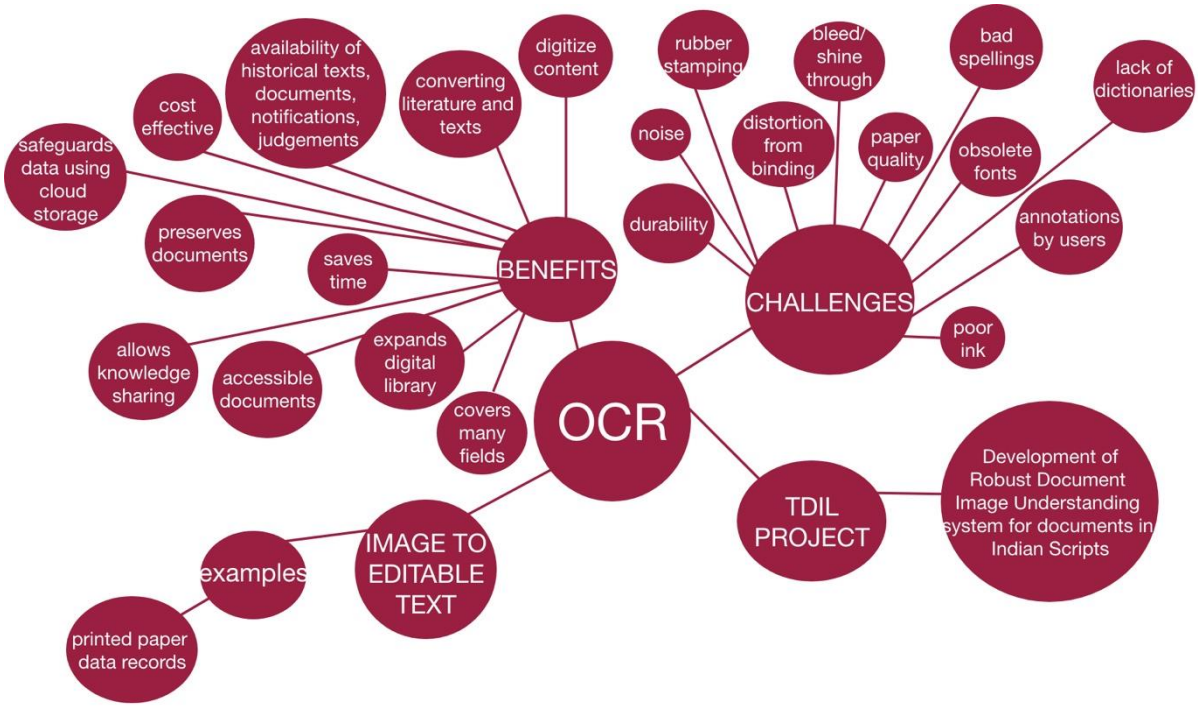


Figure 3. 2: Components of Optical Character recognition

The output of the OCR should ideally be the same as input in formatting. The Indian OCR system is simplified robust software with a reasonable performance for possible conversion of printed documents into electronically accessible documents. This system is an outcome of effort of consortium of members like C-DAC, IIIT Hyderabad, IISC Bangalore, IIT Delhi, etc. It has been developed for 13 Indian languages/scripts, namely, Assamese, Bangla, Gurumukhi, Hindi, Kannada, Malayalam, Tamil, Telugu, Urdu, Gujarati, Oriya, Manipuri, and Marathi, and has been making promising progress. Some of these functions are depicted in graphic (fig 3.2).

Introduction

The oral traditions of the Indian subcontinent are second to none in terms of the rich cultural heritage and quality literature in the form of oral traditions, and later, written texts. Oral

traditions were gradually penned down which have grown to become volumes of pan academic work. These have existed in India for the past 3000 years. These scholarly works have accumulated in silos through the past millennia and were copied and distributed till the printing press took over. The printing press used block printing to churn out more copies and ushered in a revolution that saw the end of the medieval age in western Europe. In India, scholarly work has been a continuous process and there are mountains of literature that either exists only in written, or print. This is true for all documentation before the ushering of the third and fourth industrial revolution.

The age of digitisation brings into fore front, the possibilities of making texts available in one's own native language, and data analysis and analytics on a scale and speed unimaginable in the past. Yet the documents, manuscripts or texts available are either in written or printed form and are not digitised. There has been steady progress in the attempt to digitise documents by scanning them for storage, yet this is not enough to create a central bank of documents that can be indexed and whose text can be searched as well as used for analysis. The need of Optical character recognition (OCR) and Handwritten Text Recognition (HTR) stems from the burden to utilise the possibilities of social development using ICT tools that can give unlimited knowledge access to the marginalised, backward and poor communities. Fig 3.3 depicts the complex technical process underlying OCR.

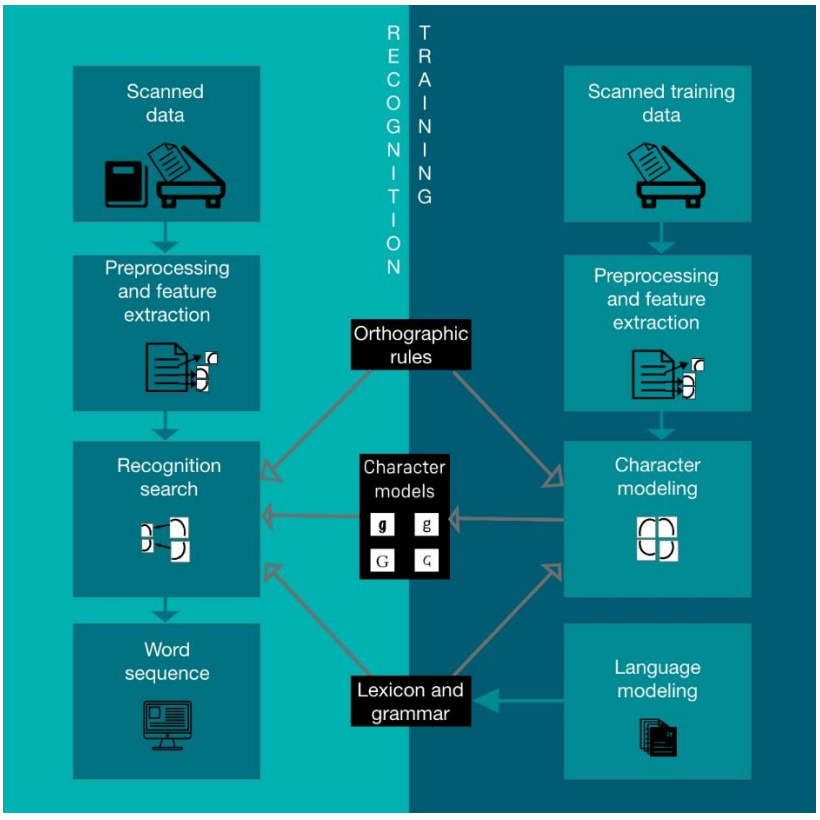


Figure 3. 3: Process of the OCR

Benefits

OCR is a utility tool for digitizing content. It has three basic components- scanning, recognition and reading text. It uses image pre-processing to clearly extract and convert text from any image of a document, into machine-encoded text using mechanical or electronic conversion. The main purpose of OCR is to digitise printed texts so that they can be electronically searched, stored more compactly, displayed on-line through virtual data rooms, and used in machine processes such as machine translation, text-to-speech and text mining. This is a crucial step in digitizing literature and texts in the country that are available in physical copies.

This has an enormous impact on the availability of historical texts, documents, notifications, judgements, etc. OCR makes such texts available for indexing, searching and sorting, saving time by making access to these documents easier as well as safeguarding the data using cloud storage. This directly translates to multitudes of data being made available for research purposes, as well as cost effective storage and transportation. The aim is to preserve these documents and make them fully accessible, searchable and processable in digital form hence reducing turnaround time for business. Knowledge contained in paper-based documents is more valuable for today's digital world when it is available in digital form. Hence, OCR is a necessity for the development and expansion of digital libraries and other knowledge networks. OCR is widely used in many fields and by many institutions including education, finance, healthcare, tourism, agriculture and governance. OCR has made countless texts available online and has been providing its services to people from all spectrums of society, saving money for students, businesses, governments, the general public and allowing knowledge to be shared.

Case Study: e-Aksharayan OCR

On 30 July 2018, a consortium led by IIT Delhi launched an OCR application that was capable of processing printed documents in seven Indian languages. Known as the e-Aksharayan Web OCR, this software is extremely useful in the movement to digitise vernacular information, as it allows a multiplicity of texts to be scanned and made available in formats that can be easily shared through the internet or other electronic media. Developed under the TDIL programme, the e-Aksharayan can either be used online or downloaded as a standalone application.

As of now, the e-Aksharayan has support for 7 Indic scripts - Bangla, Devanagari, Gurumukhi, Kannada, Malayalam, Tamil and Kannada. Functionality for 6 others - Oriya, Gujarati, Tibetan, Assamese, Manipuri and Urdu - is being tested.

The entire process of scanning documents, recognising characters and creating editable text files takes less than five minutes. e-Aksharayan comes fully equipped with noise cleaning, skew detection and binarization modules. This greatly improves the quality of text recognition, especially for documents that are slightly unclear or damaged.

The e-Aksharayan is particularly useful to those sections of Indian society that are otherwise prevented from accessing written information. WORTH Trust, a Chennai-based organisation that works with the physically disabled, has used the software to digitise and convert 600 Tamil books to the Braille script. Likewise, a website called KannadaPustaka uses the e-Aksharayan, along with TTS programmes, to create audio textbooks for blind students in Karnataka. These anecdotes illustrate how this open-source software can drastically improve people's lives at the

grassroots level.

Challenges

There are multiple challenges that are faced by OCR programs that are inherent in historical texts as well as modern prints. Paper quality that has been used is a major factor as the quality of the paper determines the durability of the text. Other factors include Distortion from binding, Bleed through / Shine through, Poor inking, Obsolete fonts, Noise and Annotations by users, Lack of dictionaries / Bad spelling, Rubber stamping, etc. These challenges are being gradually overcome with the projects that have been ongoing under the TDIL programme with the participation of consortium members.

TDIL Projects

Project details as provided by the project teams

Table 3. 1: Project details OCR

S No	Project Components	Details
1.	Title of Project: Start Date & Completion Date:	DEVELOPMENT OF ROBUST DOCUMENT IMAGE UNDERSTANDING SYSTEM DOCUMENT IN INDIAN SCRIPTS July 2010-June 2015
2.	Implementing Agencies / Participating Institutes:	IIT DELHI, C-DAC NOIDA, C-DAC PUNE,IIIT HYDERABAD,ISI KOLKATA,IISC BANGALORE, IIT BOMBAY, IIT KHARAGPUR, PUNJABI UNVERSITY, PATIALA, CENTRAL UNIVERSITY, HYDRABAD, MS UNIVERSITY, BARODA, IIT ALAHABAD
3.	Chief Investigator with Contact Detail:	PROF SANTANU CHAUDHURY PROF BREJESH LAL, IIT DELHI
4.	**Project Details:** The Objective of the project is to develop robust OCR for printed Indian scripts, which can covert scanned printed documents into machine readable/editable format (Unicode text). The implementing agency comprises of consortium members with IIT Delhi as Consortium Leader and sponsored by Ministry of Communication and Information Technology, Government of India. This system has been developed to facilitate the digitisation of the multi-lingual textual images. The OCR system can process documents in languages like Bangla, Devanagari, Malayalam, Gujarati, Telugu, Tamil, Kannada, Gurumukhi, Oriya, Tibetan, Boro, Urdu, Assamese, Marathi and Manipuri. Indian Language OCR, being	

a consortium-based project, has a hybrid approach, designed to work with platform and technology independent modules. This system has been developed to facilitate the digitisation including symbols and numerals document images having complex layout and varying font styles. OCR is developed in Windows, Linux and Web version.

Features of Present OCR Engine are:

1. The potential of OCR is enormous as it enables users to harness the power of computers to access printed documents in Indian language/scripts.
2. A number of pre-processing routines are available such as skew detection and correction, Noise removal and thresholding to convert an input grey-scale document image into clean binary image for successful recognition. Other pre-processing steps can be colour image processing, Dithering/Colour Highlight/Colour Stamp/Underline/Annotation/Marginal Noise Removal and Text-Non-Text Separation.
3. Present version of OCR supports major Indian languages/scripts- Assamese, Bangla, Gurmukhi, Hindi, Kannada, Malayalam, Tamil, Telugu, Urdu, Gujarati, Oriya, Manipuri and Marathi.
4. It converts printed document images to editable text with up to 90-95% recognition accuracy at character level & 85-90% at word level.
5. Current version of OCR takes 45 to 60 sec to process an A4 size document.

5. **Deliverables/outcome achieved in physical terms:**
The ACCURACY RATE OF DIFFERENT OCRS

Language	Accuracy
Gurumukhi	97.18
Malayalam	96.1
Marathi	92.64
Gujarati	91.51
Oriya	89.63
Hindi	93.53
Assamese	96.62
Tibetan	89.4
Manipuri	94.6
Bangla	94.54
Kannada	84.56
Tamil	80.64
Urdu	89.8

	Telugu	91.3	

3.3.2 DEVELOPMENT OF AUTOMATIC SPEECH RECOGNITION(ASR)

Overview

Automatic Speech Recognition (ASR) technology in today's age has advanced considerably and is now being used my millions of individuals to automatically create documents from dictation. ASR works by pattern matching digitised audio of spoken words against computer models (i.e., computer representations) of speech patterns to generate a text transcription. An ideal ASR system would be able to process speech audio into an error free, word-for-word transcription of the speech. The goal of an ASR system is to accurately recognise the words in speech spoken by any person in any environment, in the same way humans can. However, due to a number of factors, including the huge variations in normal human speech, perfect, verbatim transcription is not currently feasible. The advent of powerful computing devices further gives hope to this relentless pursuit, particularly in the past few years.

Introduction

Communication is the single largest factor that has accelerated the human species to form civilisations and distinguished them from others in the animal kingdom. Speech is the primary mode of communication among human beings. The number of languages that exist today are a source of fragmentation on linguistic lines and hinder communication between different linguistic groups. In the past, this communication gap has been bridged using translators and translation of texts. In the age of the fourth industrial revolution, the usage of ICT tools for communication and information sharing has tried to transcend the linguistic boundaries that limited interaction and knowledge exchange. the first step in digital translation and interaction is conversion of speech to text, to make it available for other ICT tools for further action. The mind map in Fig 3.4 accurately summarises the applications, implications and challenges involved in ASR technology.

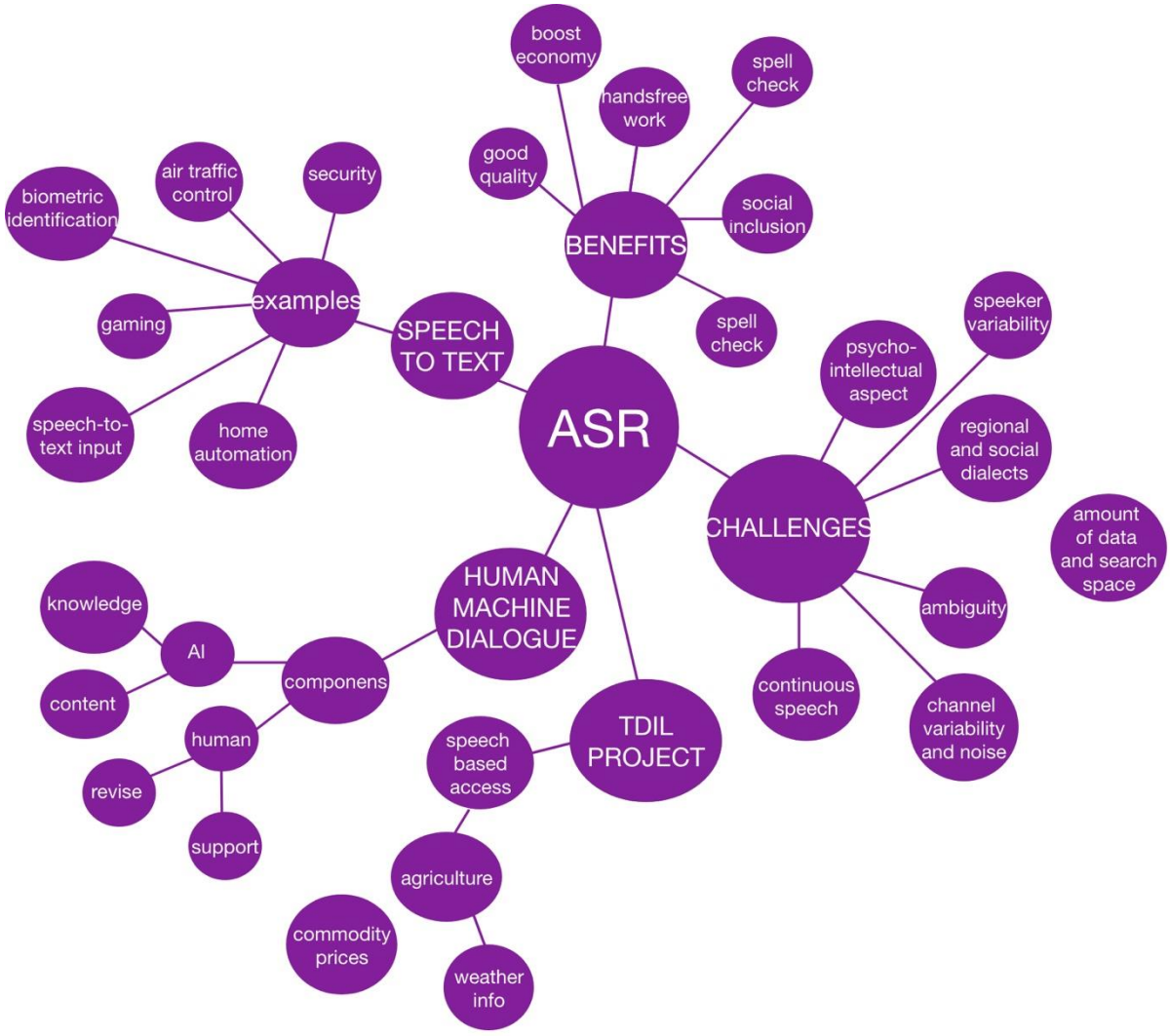


Figure 3. 4: Component of Automatic Speech Recognition

Automatic Speech Recognition (ASR) has been worked upon for the past 60 years. ASR radically changes the way transcriptions of oral dictations are taken, replacing a second person jolting down verbatim. It enables voice commands that we use daily in our smartphones for controlling applications. ASR is now also been used to assist the specially-abled in communication. It can also be used to build an extensive knowledge-based ecosystem. Applications where ASR is used, vary from simple tasks to more complex ones. Some examples of are speech-to-text input, air traffic control, security and biometric identification, gaming, and home automation.

Benefits

The benefits of ASR are enormous. The act of speaking is much faster and a more accurate expression of one's thoughts than writing for most people. As voice recognition technology is faster than typing, it offers a better and more accurate alternative to it. The quality of transcripts from ASR has been getting better and better over the recent years - even though they need to be proofed and checked for quality, the quality is fairly good. It also enables handsfree and dedicated work for professionals, enabling them to focus on their core functions without paying attention and dedicating time to the routine task of typing. This also frees the professional from focusing on their spellings as speaking directly into a digital device reduces these mistakes

considerably. ASR is the first step in a chain that is completed by converting text to speech. It is a crucial component in the processes in enabling communication between different linguistic groups and sectors. it has a direct socio-economic impact on the ability of the economically weaker section of the society everywhere, especially in India. ASR will also facilitate a new generation of economic opportunities with the advent of Big data, Neutral Networks and Artificial intelligence that work on the implications resulting from improved Text to speech programs.

Challenges

While ASR as a tool has many benefits, there are still many challenges experienced in the technology. From the psycho-intellectual aspect there are problems around human understanding of speech, spoken language and written language, multimodality in human-human communication, background noise. Other difficulties present include continuous speech, channel variability and noise, speaker variability which itself includes phone realization which is the way of speaking the same word which could result in a different pronunciation every time, making the acoustic wave of the speech vary over time for the same utterance, accent, gender of the speaker and anatomy of the vocal tract speed of speech. Other challenges are regional and social dialects, amount of data and search space, ambiguity of homophones, which are words that sound the same but have different orthography and word boundary ambiguity.

TDIL Project

Table 3. 2: Project details ASR

S No	Project Components	Details
1.	Title of Project: Start Date & Completion Date:	SPEECH-BASED ACCESS OF AGRICULTURAL COMMODITY PRICES AND WEATHER INFORMATION IN 11 INDIAN LANGUAGES/DIALECTS Phase 1: May 2010 - February 2013 (6 Languages) Phase 2: February 2013 - October 2018 (5 Languages)
2.	Implementing Agencies / Participating Institutes:	IIT MADRAS, IIT BOMBAY, IIT GUWAHATI, IIIT HYDERABAD, TIFR MUMBAI, CDAC KOLKATA, IIT KHARAGPUR, IIT BHUBANESWAR, SIT TUMKUR, DA-IICT AHMEDABAD, BIT MESRA
3.	Chief Investigator	Dr S Umesh, Professor, Dept of Electrical Engineering, IIT Madras

5.	Project Details: Nature of the Project: Application-oriented Research, Design and Development (R&D) having production potential. The automatic speech-based price retrieval system, is built based on the following facts: • Deep penetration of mobile phones in India's rural areas enables the use of mobiles as a medium of getting information. • Very limited internet connectivity in India in rural areas and therefore web-based services may not be suitable. • Very limited reading and writing skills of the majority of target audience which means speech is the only means of communication for them. The main task was to develop an interface using relevant speech technologies so that even the most novice of users was able to get the relevant information with minimum human intervention. The speech interface was developed separately for each of the eleven languages, taking into account the needs for that language.
6.	Deliverables/outcome achieved in physical terms: a) Built Speech based systems for the following states: Andhra Pradesh, Assam, Bihar, Gujarat, Jharkhand, Karnataka, Maharashtra, Odisha, Tamil Nadu, Telangana, Uttar Pradesh and West Bengal. b) Standardisation of ILSL12 Pronunciation Dictionary across different languages. Effort has been made to standardise the phone set used across the different languages. c) For each of the eleven languages, the speech data was collected from farmers (end users) across different districts to capture the dialectical variations. It includes: • Spoken data from about 1000 farmers, in each language and the corresponding verbatim transcription. • Agricultural commodity names, district/mandi names requests in each of the eleven Indian languages/dialects along with associated pronunciation variants/accents. • Weather forecasting information, speech data of farmers, recorded in the natural environment, using the farmers own handset to capture environment and mobile network variabilities. • Data also contains, response to open-ended questions, thus capturing natural variations in free-flow speech, such as hesitation, pause, breath-noise, etc.

Human–Machine Dialogue
A true dialogue goes beyond an interrogation for information (either from the user to the machine or vice versa). A dialogue involves response based on the previous state of the

conversation. Both the user and the machine should be able to ask for clarification or extension to the previous information exchange, or to initiate a new domain of interaction. This more natural form of dialogue is referred to as a mixed initiative or variable initiative dialogue

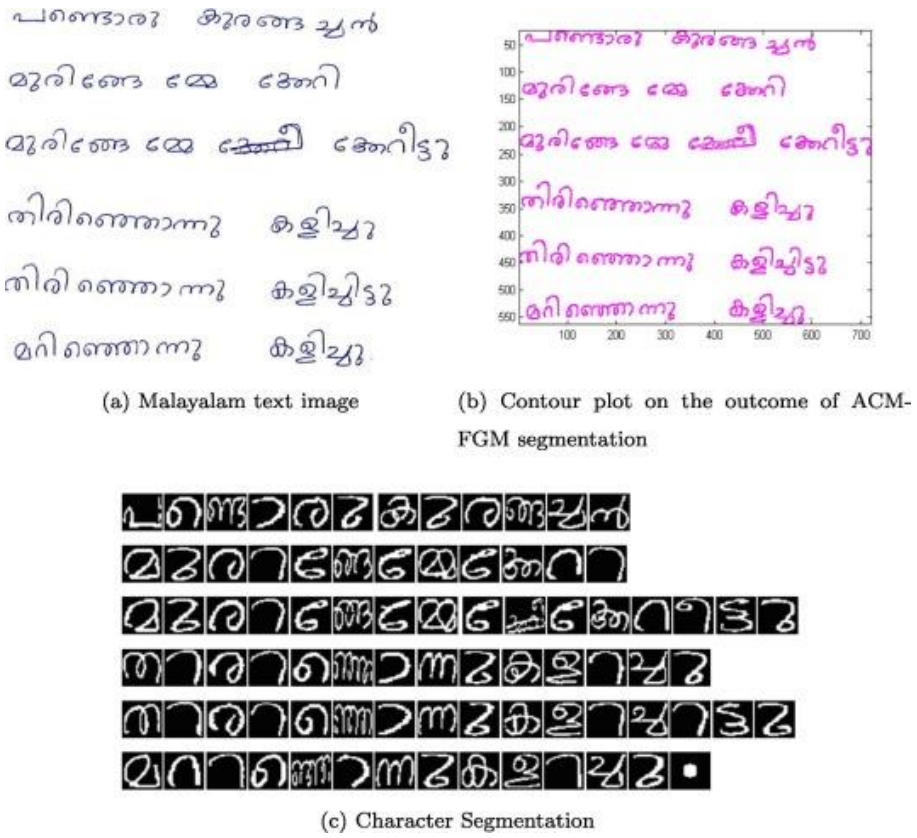
Research issues in this area can be divided into two broad categories. One involves artificial intelligence because, in order for a machine to be able to perform natural dialogue, it must have acquired an ability to communicate (as opposed to just recognise). Such ability is far beyond speech recognition and understanding; it must have a store of "knowledge" and "content," not just the protocol or handling of the communication mechanism, to support the communication. Knowledge representation and retrieval, database organisation and search, semantic inference, and decision support are all required to various degrees. This is obviously a very broad and long-term challenge.

The other research need is much narrower but immediate. Before it is possible to design a machine that can communicate, it is often desirable to provide system design tools to allow human intervention, either at the application design stage or during run-time. Tools that allow the human–machine interaction system designer to develop a task based on his or her anticipation or envisage of the system behaviour in response to a majority of users are very useful. For example, tools that provide efficient design of dialogue states & flow and allow the developer to revise and support the operation of the system are deemed extremely valuable.

3.3.3 DEVELOPMENT OF ON-LINE HANDWRITING RECOGNITION SYSTEM (OHWR)

Overview

The Online Handwriting Recognition (OHWR) is a system that involves a process where handwritten messages can be recognised by processing the received data. It is the process of converting handwritten characters into machine format. It operates using different algorithms, such as the ACM-FGM algorithm (shown in Fig 3.5).



(a) Malayalam text image

(b) Contour plot on the outcome of ACM-FGM segmentation

(c) Character Segmentation

Figure 3. 5: Character segmentation based on ACM-FGM algorithm

Data entry using a pen forms a natural, convenient interface. The large number of writing styles and the variability between them makes the problem of writer-independent unconstrained handwriting recognition a very challenging pattern recognition problem. Compared to the research in 1980s, the research efforts in the 1990s aimed to further relax the constraints of handwriting, namely, the adherence to standard stroke orders, stroke numbers and the restriction of recognition to isolated characters only.

Introduction

The Sumerian archaic (pre-cuneiform) writing and the Egyptian hieroglyphs are generally considered the earliest true writing systems, both emerging out of their ancestral proto-literate symbol systems from 3400–3100 BC. From then on, language has evolved exponentially, with innumerable writing systems rising and falling out of use between then and the 21st century. The printing press and, more recently, the typewriter were able to reduce dependence on handwritten texts meant for storage, record keeping and official communications, yet most of literature and texts was being jotted down by hand. Today, much of all the content is being created in digital form by typing, yet most of the past records, literature and data is stored in hard copies and are many a times handwritten. Even though we have made extensive strides in the field of Optical Character Recognition (OCR), it is not meant for processing handwritten text. Fig 3.6 is a mind map detailing the features, benefits and challenges of OHWR technology.

Figure 3. 6: Components of Online-Handwriting Character Recognition

The need for Online Handwriting Recognition (OHWR) has been recognised and was first worked on by an IISC Bangalore based team in 2006. OHWR is used to recognise handwritten texts and convert the scanned document into editable text. This opens a lot of avenues for research in various fields such as online recognition, signature verification, Postal address interpretation, Bank Check processing, writer recognition, etc.

Benefits

The Online Handwriting Recognition, in comparison to offline recognition, focuses on the work where recognition needs to be performed simultaneously as the time of writing. This, thus, requires specialised tools such as a digitizing tablet to capture the strokes of the pen as they are being worked upon. This trace from the writer's pen is stored and can be later be used to create a static image of writing that facilitates the application of offline character techniques. The accuracy levels of online recognition are also greater than the offline one. Further, the data requirement for the online recognition is much smaller as compared to the offline one.

Additionally, the OHWR system can be more adaptive where the system gives a prompt feedback that can be used to give further training to the recogniser. This is also a real time process that picks up the dynamic information of writing. The system requires very little processing and eases the operation of segmentation. Online handwriting recognition system also minimises ambiguity as the information from the pen trajectory can be used to assist ambiguous optical characters. The recent transition in the field of personal computing from the desktop to handheld devices has been made possible by the paradigms suited for hand entry has also been made possible by the various developments in the online handwriting recognition systems.

Challenges

The OHWR System faces various challenges that hinder the employment of the system to its full potential. One of the major problems that one comes across when dealing with any handwritten script is that the characters within it, written by different persons representing the same character, are not identical and can vary in both shape and size. The variations in different writing styles, different stroke orders, and efficient data presentation may also obstruct adept working of the system. This is also true for Indian Scripts that have various handwriting styles that vary according to the region, which can make recognition a challenging task. The similarities of distinct character shapes and the ambiguous writing further complicate the recognition process.

Another challenge one encounters when dealing with OHWR is that it requires the use of specialised equipment like a stylus to work and cannot be applied on written or printed documents.

TDIL Projects

Table 3. 3: Project details OHWR

S No	Project Components	Details
1.	Title of Project: Start & Completion Dates:	DEVELOPMENT OF ONLINE HANDWRITING RECOGNITION SYSTEM FOR INDIAN LANGUAGE (OHWR) – PHASE II –DEPLOYMENT OF AN APPLICATION AND IMPROVEMENT OF ENGINE PERFORMANCE. 31st December 2011 and June 2016
2.	Implementing Agencies/ Participating Institutes:	IISC BANGALORE; IIT MADRAS; IIIT HYDERABAD; IIT GUWAHATI; CDAC PUNE; ISI KOLKATA; THAPAR UNIVERSITY PATIALA.
3.	Consortium Leader with Contact Details	Prof. A G Ramakrishnan Dept. of Electrical Engineering Indian Institute of Science, Bangalore
4.	Project Details: 1. Pilot deploy a prototype application in languages where the recognition (test) performance is high, to demonstrate the effectiveness of the technology and improve the applications from the point of view of actual users. 2. Obtain a character (Unicode) recognition performance of 95% for carefully-written data and 90% for casually written data, for the following scripts: • Tamil • Telugu • Kannada • Malayalam	

	• Devanagari • Bangla 3. Add new development partners for Gurmukhi and Assamese, and create recognition engines in Assamese and Punjabi, after collecting database in these languages matching the level of collection of other languages in phase I and achieve performance of 90% for carefully written words. 4. Collect additional akshara level as well continuous sentence level handwritten data. 5. Explore the new tablet-based devices in the market and choose a promising device for application development. 6. Develop recognition at the level of handwritten sentences, by introducing word level segmentation. 7. Create and make available, a web demo of the well performing OHR recognition engines, in order to popularise the technology and facilitate trials by interested people, as well as get feedback for performance enhancement. 8. Explore possible applications for the handwriting recognition technology on handheld devices, by implementing good engines developed in phase I, after optimisation.	
	Deliverables 1. Generic online Inputting tools for Android and Windows for all the languages. 2. Census Data Collection application for Android platform for all languages. 3. Smart Input Panel for all languages in Windows 7 or higher Tablet PC with stylus (sentence and paragraph level handwritten recognition). 4. Bigram language model-based recognition. 5. Showcasing technology by integrating Tamil and Hindi OHWR engines with e-gov applications including Census data collection application.	Achievements 1. Developed and uploaded 2. Developed for 2 languages (Hindi and Tamil). 3. Developed & Published for Hindi and Tamil. 4. Developed for CDC

3.3.4 MACHINE TRANSLATION SYSTEM (MT)

Overview

This component aims to cover the analysis of the primary research of the TDIL programme by discussing the benefits and challenges of Machine Translation.

The research further explains Machine Translation in a two-pronged approach - Translation from English to Indian languages and translation of Indian to other Indian languages. In Translation from English to Indian languages, it covers two projects called AnglaMT System and Anuvadaksh System. Even under Indian to Indian language Machine Translation, Sampark System has been

discussed. This component aims to cover the analysis of the primary research of the TDIL programme by discussing the benefits and challenges of Machine Translation (depicted in the mind map in Fig 3.7).

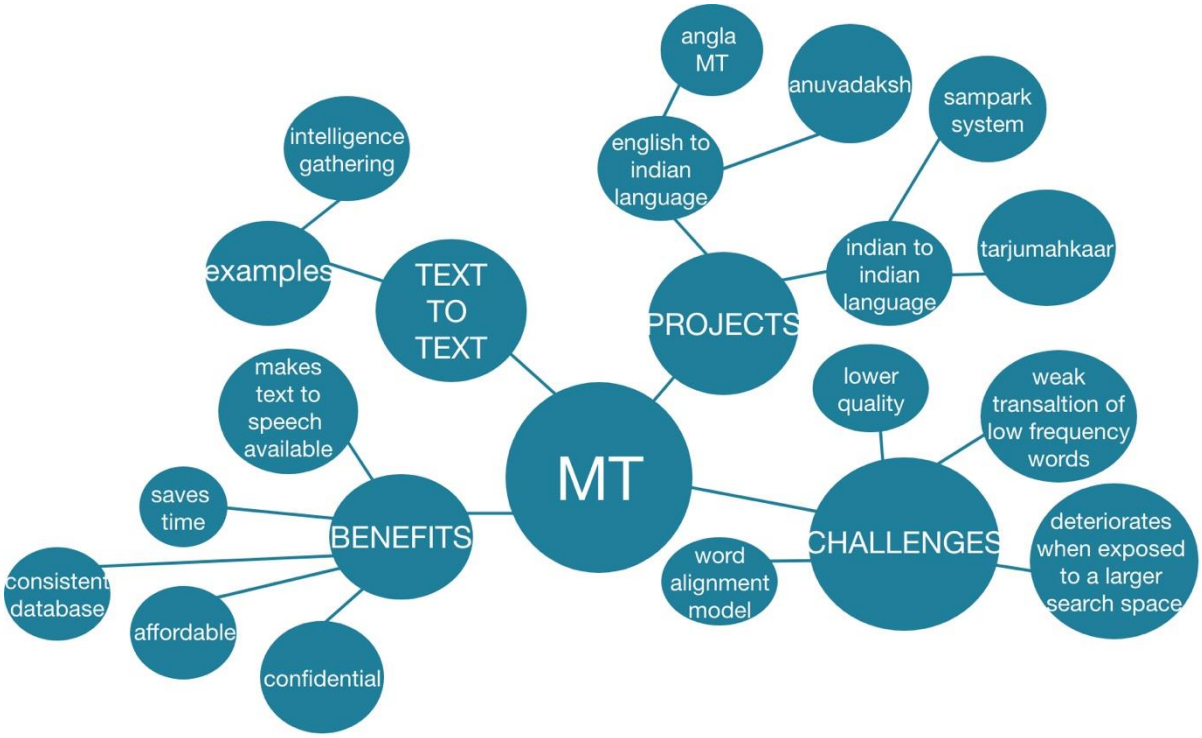


Figure 3. 7: Components of Machine Translation

Development of Sanskrit Hindi Machine Translation System (SHMT), which can also be considered as a part of Machine Translation, is also discussed separately under this topic of Machine Translation to provide it a holistic view.

Introduction

There are 22 constitutionally recognised official languages and several hundred other languages in India. This is beside 122 major languages and 1599 other languages. Under the aegis of our former Prime Minister, P.V.Narasimha Rao, a fully dedicated centre for translation studies was established in University of Hyderabad, but much before that, Centre for English and Foreign Language (CIEFL) was established in Hyderabad with regional centres like Lucknow and Shillong, to promote the aspect of translation. Translation is required in order to pass the work of state government to the central government and aid in communication in other fields as well. Due to this diversity in languages, it is almost impossible for everyone to know all the languages. Even the newspapers are published in various languages. So, to overcome all these limitations, translation is required, and translation is the obvious go-to for abolishing the communication barrier. MT is a subfield of computational linguistics, sometimes also referred to as Natural Language Processing (NLP), and employs multiple data assets to build phase models which are used to translate text. Machine translation (MT) can be defined as an automated system that analyses text from a source language (SL), by applying some computation on that input, and

produces equivalent text in a required target language (TL), ideally, without any kind of human intervention.

The most immediate division of translation purposes involves information acquisition versus dissemination. The classic example of the former purpose is intelligence-gathering: with masses of data to sift through, there is no time, money, or incentive to carefully translate every document by normal (i.e., human) means. Scientists, more generally, are faced with this dilemma: there is already more to read than can be read in the time available, and having to labour through texts written in foreign languages - when the probability is low that any given text is of real interest - is not worth the effort.

Benefits

Machine Translation is a crucial element in the pursuit of realising the dream of perfect, human speech to speech translation. It is the technology that processes and translates text from one language to another and makes it available for text-to-speech and other similar projects that are crucial steps in spreading the benefit of this technology. The average human translator can translate around 2,000 words a day; it is evident that the rate of machine translation is exponentially faster than that of human translation. Hence, using the machine translation system enables you to save your time while translating large texts. It is also comparatively cheaper, although initially it might look like an unnecessary investment, but in the long run it is a very small cost considering the return it provides. Machine translation can memorise key terms and phrases that are used within a given industry. Hence, they have an ever-increasing consistent database, which leads to translations that are very consistent across the entire file, something that is more difficult to achieve when using multiple human translators. Confidentiality is another matter which makes machine translation favourable.

Case Study: MANTRA Rajbhasha

The translation of legal and administrative proceedings is a high priority in the language technology industry. Whether it is in the United Nations or the European Union, there is always a demand for both translators as well as computational linguists to facilitate communication between speakers of different languages. This is because of the sheer level of linguistic diversity within both entities. In India, a country with 22 national languages and many hundred dialects, these linguistic experts are key to maintaining national integrity. They are especially needed to facilitate the smooth functioning of the government. This case study examines MANTRA Rajbhasha, a tool that translates the proceedings of the Rajya Sabha from English to Hindi.

The Rajya Sabha is currently equipped with interpreters for all 22 of the country's national languages, making it possible for members of the House to make statements in their native languages. However, the setup would be incomplete without a mechanism to translate the various documents that are generated, tabled and discussed in the House's proceedings - and this is where MANTRA comes in.

MANTRA, or the Machine-Aided TRAnslation Tool, is a program developed by CDAC Pune under TDIL programme. In 2007, CDAC Pune released a standalone version of the program, known as MANTRA Rajbhasha, that had been customised to translate the proceedings of the Rajya Sabha. This would allow the representatives from different Indian states to easily access the documents

officialised under parliamentary convention.

MANTRA Rajbhasha handles all of the House's paperwork, including Papers to be Laid on the Table (PLOT), List of Business (LOB), Bulletin-I and Bulletin-II. Initially filed in English, these documents are translated and made available in Hindi, which is understood by roughly two thirds of India. The software takes about one minute to translate each document and provides an accuracy of 90-95% in its translations (although minor edits often have to be made by human translators).

While both English and Hindi are official languages in the Indian Union, English is seen as a hangover from British colonialism. Moreover, it is spoken fluently by just 5% of Indians. Hindi, on the other hand, is understood by over half the country. Moreover, it is projected as a unifying force by the Indian government. MANTRA Rajbhasha thus defies India's imperial heritage and brings government proceedings to its servants in a language that is seen as quintessentially Indian.

Challenges

- Neural Machine Trasnslation systems have lower quality out of domain, to the point that they completely sacrifice adequacy for the sake of fluency.
- NMT systems have a steeper learning curve with respect to the amount of training data, resulting in low quality in low-resource settings, but better performance in high resource settings.
- NMT systems that operate at the sub-word level (e.g. with byte-pair encoding) perform better than Statistical Machine Translation systems on extremely low frequency words, but still show weakness in translating low-frequency words belonging to highly inflected categories (e.g., verbs).
- NMT systems have lower translation quality on very long sentences but do comparably better up to a sentence length of about 60 words.
- The attention model for NMT does not always fulfil the role of a word alignment model but may in fact dramatically diverge.

In India, work has been done in Machine Translation in a two-pronged approach. Translation from English to Indian languages and translation of Indian to other Indian languages.

English to Indian Language machine Translation System (EILMT)

There are two projects under the English to Indian Language Machine Translation programme:

AnglaMT

URL: http://tdil-dc.in/mt/common.php

It is a Rule Based Machine Translation System, designed for translating Text in English to Indian languages with pseudo-interlingua approach. It analyses English only once and creates an

intermediate structure with most of the disambiguation performed and is used to generate Indian Language translated output. This approach was adapted to create eight MT systems.

Features

- System made available for public usage on www.tdil-dc.in
- Translated output represented in script of the target language, Devanagari and Roman scripts
- Input through keyboard or from file.
- Successful adaptation of AnglaHindi to Punjabi, Bangla, Urdu, Telugu, Malayalam, Assamese and Nepali languages.
- On-screen keyboard facility for post editing by user
- Translation service available through SMS also

Table 3. 4: Project details MT (AnglaMT)

S No	Project Components	Details
1.	Title of Project/Technology: Start Date & Completion Date:	DEVELOPMENT OF ENGLISH TO INDIAN LANGUAGE MACHINE TRANSLATION SYSTEM BASED ON ANGLABHARATI TECHNOLOGY 23 February 2011 to 21 November 2015
2.	Implementing Agencies / Participating Institutes	CDAC- NOIDA, KOLKATA, HYDERABAD, & TRIVANDRUM
3.	Chief Investigator with Contact Detail	1. KARUNESH KUMAR ARORA(CDAC NOIDA) 2. MR. BHADRAN V K (CDAC TRIVANDRUM) 3. MR. SANJAY CHOUDHURY(CDAC KOLKATA) 4. MS. RADHIKA KAMBHAM(CDAC HYDERABAD)
4.	Project Details: This project was based on a rule-based machine translation system for translating text in English to Indian languages (Assamese, Bangla, Hindi, Malayalam, Nepali, Punjabi, Telugu and Urdu) developed in Consortia mode. The consortium included member institutes CDAC Noida, Kolkata, Hyderabad and Trivandrum. The aim was to use AnglaMT technique developed in earlier phase and to improve translation system and to extend the technology to include newer language. It exploits commonality among Indian languages and addresses syntactic and semantic divergence between English and Indian languages.	

5.	**Deliverables/outcome achieved in physical terms:** 1. **MT Systems deployed at https://tdil-dc.in, https://anglamt.tdil-dc.gov.in (Meghraj cloud).** 2. **Various resources and tools are made available for research and development community.** 3. **Multi-word Expressions / Named Entities developed and lexical database generated for Health & Tourism domains in the development duration.** 4. **Both Bilingual & Monolingual corpus developed for various Indian languages.** 5. **Various sub-systems/modules were developed and can be used by researchers for developing other systems including morphological analyser, morphological synthesisers etc.** 6. **Various support tools like Lexical Database Management System, Admin Interface and Translator's Workbench were developed for managing, monitoring and control.** 7. **System's capability successfully demonstrated through translating books of State Institute of Languages, Govt. of Kerala and webpages of Vikaspedia portal.** 8. **Various Institutes/organisations have shown interest for utilisation of services and commercialisation of technology**

Anuvadaksh System

URL- http://tdil-dc.in/eocr/index.html

Anuvadaksh system, which is Multi-engine, multi-output system, includes three engines, viz., TAG, SMT & EBMT, and targets the following language pairs:

English to Hindi	English to Marathi	English to Bangla
English to Urdu	English to Tamil	English to Odia
English to Gujarati	English to Bodo	

It allows translating the text from English to the preceding eight Indian. Domains covered for translation through the system are Tourism, Health care and Agriculture.

Following are the features of the ANUVADAKSH system on C-DAC Portal:

- Tree Adjoining Grammar (TAG) engine facilitates the translation for all the eight language pairs.
- Non-Login usage facility
- Language and domain selection facility
- User can copy–paste a text and get instant translation
- Pre-Processing module prepares input text into engine suitable form
    - Morphological Analyser
    - Part of Speech Tagger
    - Named Entity Recogniser
    - Chunking and Clause Identification
- Post-processing module provides additional features for EILMT Translation engine

- Morph Synthesiser for smoothening the translated output
- Multiple translation options
- Synonym selection option
- Transliteration Facility

Indian to Indian language Machine Translation (ILMT):

The project under the Indian to Indian Language Machine Translation programme:

Sampark System: Automated Translation among Indian Languages

URL: https://tdil-dc.in/components/com_mtsystem/CommonUI/AdvaceFeature.php

Sampark is a multipart Machine Translation system ILMT that has developed language technology for 9 languages resulting in MT for 18 language pairs. Sampark uses Computational Paninian Grammar (CPG) approach for analysing language and combines it with machine learning. Thus, it uses both traditional rules-based and dictionary-based algorithms with statistical machine learning.

14 bi-directional pairs between: Hindi with Urdu / Punjabi / Telugu / Bengali / Tamil / Marathi / Kannada

4 bidirectional between: Tamil with Malayalam / Telugu.

3.3.5 DEVELOPMENT OF SANSKRIT HINDI MACHINE TRANSLATION SYSTEM (SHMT)

Sanskrit to Hindi MT system is being developed under a project by the consortium of Hyderabad University and Jawaharlal Nehru University along with other institutions[1]. It focuses on the domain of children's stories, building multimedia and e-learning content for kids. The system will be updated by incorporating modules like Word Sense Disambiguation Module, Anaphora Resolution Module and Default Prose order generator, and will be extended for domains like Yoga and Ayurveda. Under one of the consortia, the Sanskrit to Hindi system will develop and deploy a unidirectional machine translation.
Three systems were proposed for the purpose:

- Sanskrit-Hindi Anusaarka
- Sanskrit-Hindi with IIIT parser- uses Dashboard
- Sanskrit-Hindi with UoH parser

As the deliverable of San2Hin, two systems SHMT Anusaarka and SHMT with IIIT parser were installed by consortia at C-DAC-GIST servers for testing for 100 sentences. SHMT with IIT parser is a Desktop based system whereas Sanskrit-Hindi Anusaarka system is a Web-based system. Both the systems support classical Sanskrit and Sandhi spitted Sanskrit sentences. It, however, fails to handle the context of the inputs given to the system.
A Sanskrit to Hindi Dictionary has also been developed by Sanskrit Centre of JNU.

[1] https://tdil-dc.in/index.php?option=com_vertical&parentid=87&lang=en

3.3.6 CROSS LINGUAL INFORMATION ACCESS

Understanding CLIA

With the tremendous growth of digital and online information repositories, new opportunities and new problems are created for achieving information retrieval across different languages. Online documents are available internationally in many different languages. Cross Lingual Information Access (CLIA) systems makes it possible for users to directly access sources of information which may be available in languages other than the language of query.

Cross-language information retrieval enables users to enter queries in their languages and uses language translation methods to retrieve documents originally created in other languages. CLIA is an extension of the Cross-Language Information Retrieval paradigm. Users who are unfamiliar with the language of documents retrieved are often unable to obtain relevant information from these documents. The objective of CLIA is to introduce additional post retrieval processing to enable users make sense of these retrieved documents. This additional processing may take the form of machine translation of snippets, summarisation and subsequent translation of summaries and/or information extraction. Fig. 3.8 is a mind map detailing the components of CLIA.

Figure 3. 8: Components of Cross Lingual Information Access

Need of CLIA

- Great language diversity
- Low comfort level with English
- Need for critical information in large quantity and high quality especially in agriculture, health, tourism, education, etc.

CLIA Project at TDIL

CLIA Project Started in year 2006 with tourism and health as the focus domains. The mission mode project is funded by the Government of India, Ministry of Communications and Information Technology, and Department of Information Technology (now MeitY). The CLIA portal for Indian Languages project is executed by a collaboration of academic, research institutions and industry partners, namely, IIT Bombay, IIT Kharagpur, IIIT Hyderabad, AU-KBC Chennai, AU-CEG Chennai, ISI Kolkata, Jadavpur University Kolkata, C-DAC Pune, C-DAC Noida, Utkal University Bhubaneswar and STDC, MeitY, New Delhi.

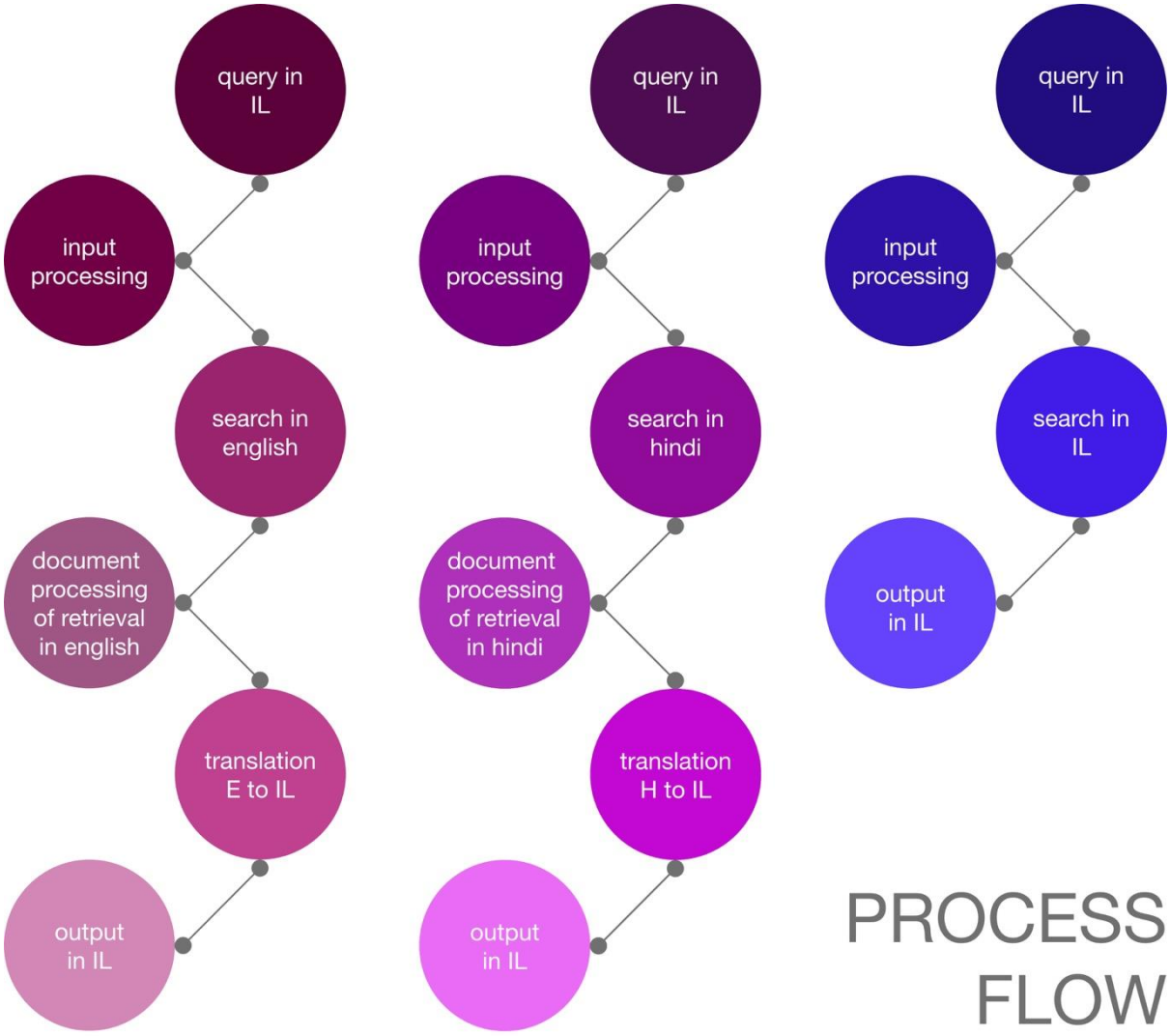


Figure 3. 9: Process flow of CLIA

The CLIA system has been developed based on the basic architecture of Nutch1, which use the architecture of Lucene2. Nutch is an open source search engine, which supports only a monolingual search. Various new or modified features of the CLIA system have been added or modified into the Nutch architecture. The main feature of CLIA is the cross lingual search, which needs the query translation, snippet translation and language independent output generation module, such as Snippet Generation and Summary Generation. The process flow within the CLIA paradigm is depicted in Fig 3.9.

In this portal, a user can submit a query in one Indian language and be able to access documents in the same language as that of the query and in Hindi and English, as well. There is also provision for translated snippet and summary in this system. The languages which are involved are Bengali, Hindi, Marathi, Punjabi, Tamil, Telugu, Assamese, Oriya and Gujarati. A user can submit a Natural Language query in a source language, further allowing them to access the documents available in the public domain in one's own language. It is heavily based on the Machine Translation System, a common example that can be cited is that of MANTRA.

Working of CLIA Portal

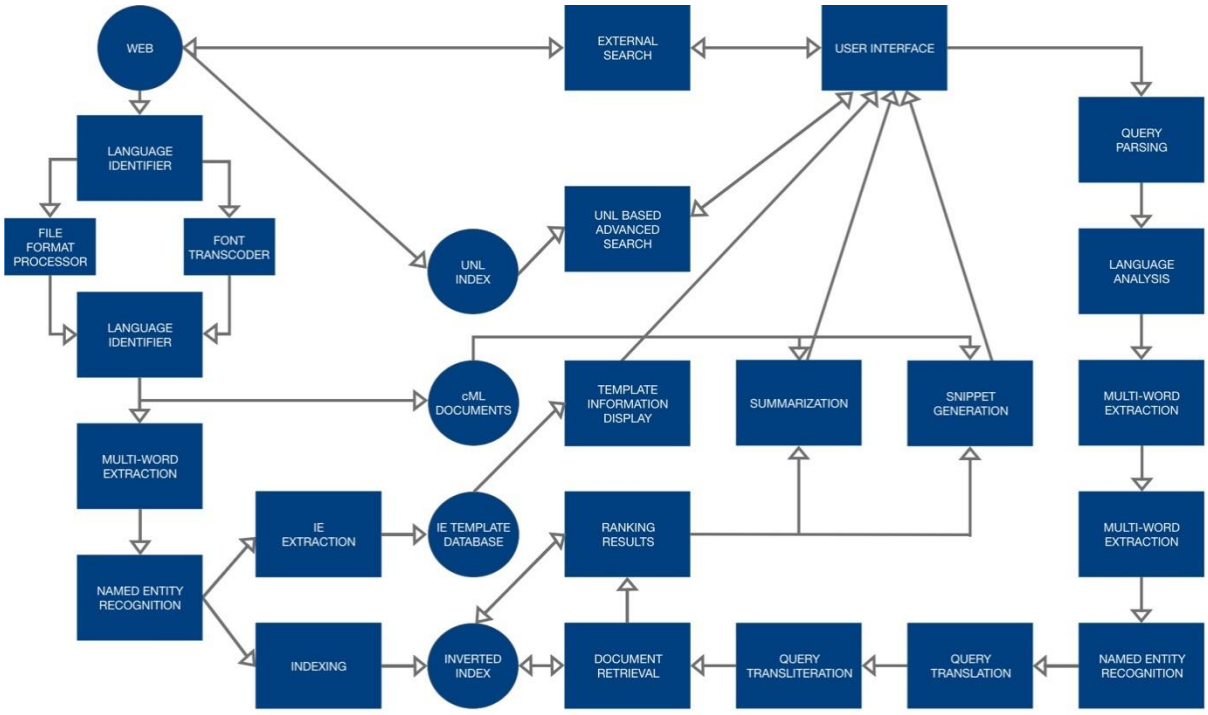


Figure 3. 10: Internal working of CLIA

The system is intended to search different documents in Indian languages. Once the user starts the system, an initial screen with logo is displayed. By default, the screen is displayed in Hindi or English, depending on the default language selected on the browser of the user. If the user wants to display this initial screen in any other language, he/she can select the language from the bottom of the screen. The mechanism by which data is accessed from different databases via the CLIA portal can be seen in Fig 3.10. The screen is then displayed in the selected language. At present, the screen is available in seven languages - Hindi, English, Marathi, Punjabi, Bengali, Tamil and Telugu. To search a document, the first activity the user performs is the selection of the source language. Selection of source language allows the user to enter the text in the

selected language. An example of a page displaying the results on the CLIA platform can be seen in Fig 3.11.

- Selection of the Source Language: The user can select the source language by clicking .....The system displays the languages available to select the source language.
- Entering String for Search: The user enters the query string on which the search is to be made in the appropriate place. The system allows the user to enter the string in the source language selected by the user using a soft keyboard for the language.
- Search the Web or the Site: Once the string is entered, the user should select whether to search the local site or the World Wide Web. The user can then click the search option to search the site for the string entered.
- Displaying the Results: Once the query is properly expanded and translated, it is used to search the web, or the local site and the documents are retrieved according to the query. The snippets of the retrieved documents are displayed in the original language of the document as well as in the source language selected by the user. Thus, if the source language selected is Bengali, the user can enter query string in Bengali, the CLIA system searches for documents in English, Hindi and Bengali either from the web or the local site. The snippets of the retrieved documents are displayed in English/ Hindi and Bengali.

Figure 3. 11: CLIA testing

- Advanced Search: The user can also select the advanced search option and the CLIA system displays all the options accordingly. The user can select here the domain in which he/she wants to search the documents. At present, the tourism and health domains are available. The user can also select the number of items to be displayed on a single page. By default, the system displays 10 items on a single page. Once the selection is made, the user can click the 'search' option to start the search. In the advanced search option, the CLIA system provides summary as well as extracted information in the form of predefined information extraction templates, of the retrieved documents along with the generated snippet. The summary and the extracted information templates can be displayed in the original language of the document as well as in the source language selected by the user.

TDIL Projects

Project details as provided by the project teams

Table 3. 5: Project Details CLIA

S No	Project Components	Details
1.	Title of Project: Start Date & Completion Date:	DEVELOPMENT OF A CROSS LINGUAL INFORMATION ACCESS SYSTEM(CLIA PHASE II) 1st September 2010 to 31st March 2017
2.	Implementing Agencies / Participating Institutes:	Implementing agencies: IIT BOMBAY Participating Institutes: CDAC, NOIDA, IIT KHARAGPUR, AUKBC CHENNAI, AUCEG, CHENNAI, CDAC PUNE, DAIICT GANDHINAGAR, IIIT HYDERBAD, ISI KOLKATA, JADAVPUR UNIVERSITY, IIIT BHUBHNESHWAR, GAWAHATI UNIVERSITY
3.	Chief Investigator:	Prof. Pushpak Bhattacharyya, CSE Department, IIT Bombay
4.	Project Details: Cross Lingual Information Access (CLIA): CLIA is a mission mode project being executed by a consortium of academic and research institutions and industry partners, and funded by TDIL, Ministry of Information Technology, Government of India. CLIA was started on 29th August, 2006, with the aim of providing a Search Engine. A large amount of information in the form of text, audio, video and other documents is available on the web. Users should be able to find relevant information in these documents. Information Retrieval (IR) refers to the task of searching relevant documents and information from the contents of a data set such as the World Wide Web (WWW). A web search engine is an IR system that is designed to search for information on the World Wide Web. There are various components involved in information retrieval. IR system has following components: • Crawling: Documents from web are fetched and stored. • Indexing: An index of the fetched documents is created. • Query: Input from the user. • Ranking: The systems produces a list of documents, ranked according to their relevance to the query. Information on the web is growing in various forms and languages. Though English dominated the web initially, now less than half the documents on the web are in English. The popularity of internet and availability of networked information sources have led to a strong demand for Cross Lingual Information Retrieval (CLIR) systems. Cross-Lingual Information Retrieval (CLIR) refers to the retrieval of documents that are in a language different from the one in which the query is	

	expressed. This allows users to search document collections in multiple languages and retrieve relevant information in a form that is useful to them, even when they have little or no linguistic competence in the target languages. Cross lingual information retrieval is important for countries like India where very large fraction of people are not conversant with English and thus don't have access to the vast store of information on the web.
5.	Deliverables/outcome achieved in physical terms: The developed Cross Lingual Information Access system (Sandhan) has a website and also a mobile application for Android for easy access of the search engine from any Android phone. System was successfully presented at IJCNLP'08 and subsequently at ELITEX'08

3.3.7 DEVELOPMENT OF TEXT-TO-SPEECH SYSTEM FOR INDIAN LANGUAGES (TTS)

Overview

The applications developed under Text-to-Speech project tries to keep the social cause as its prime objective including other initiatives such as Browser Plug-ins, SMS Reader and website to check the quality of the system. The Text-to-Speech software has been developed using Open Source FESTIVOX Framework and State-of-the-art HTS based engine into 13 Indian Languages by various prime Indian Institutes of excellence.

The research explains vision and objectives of Development of Text-to-Speech System for Indian Languages which is majorly focused upon enabling citizens to take upon the benefits of the ICT in order for them to empower moving beyond language barrier and capabilities. It further provides list of projects taken up under Text-to-Speech System.

Introduction

TDIL programme visions to empower citizens from all classes and background to take on the benefits of ICT and knowledge sharing. When we talk about empowering everyone, the various divisions of the society aren't only based on the region and class, but also on the individual abilities. With the aim of delivering services to every individual despite of their abilities and disabilities, the Text-to-Speech software was developed. The software delivers machine readable texts and figures into human voice. Thus, once the software is integrated with the screen reader it shall allow people with visual impairments and reading disabilities to listen to written works on a computer, mobile device or a tablet.

The software has been developed for 13 Indian Languages, namely, Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Gujarati, Odia, Assamese, Manipuri, Kannada, Bodo and Rajasthani, under the leadership of IIT Madras and other major institutions. IIIT Hyderabad, IIT Kharagpur, IISc Bangalore, IIT Guwahati, IIT Mandi, CDAC Mumbai, CDAC Thiruvananthapuram, CDAC Kolkata, SSNCE Chennai, DA-IICT Gujarat and PESIT Bangalore were the other major contributing institutions funded by TDIL Programme.

Vision and Objectives

The vision is to enable citizens to take upon the benefits of the ICT in order for them to empower moving beyond language barrier and capabilities. With the growing penetration of technology in the society, the future is said to be "online". India being a multilingual country, wherein the English-speaking population is between 5-7%, struggles with utilising full advantages of the Information and Communication Technology. While the ICT sector flourishes and comes up with new advancements and policies every day, a major portion of the population is deprived of such progress. More than the concept of affordability and accessibility, the adaptability of the information becomes a barrier, specifically in terms of the language used. Major, or to say almost work done in the sector of IT is in English. When we say the vision is to empower citizens at the grassroot level, the need is to eradicate the language barrier by providing a free flow and easy communication of information between both the parties. The means of communication shall be through online platforms. However, the first step to such a world is by the means of simplifying these online platforms for the citizens itself. In order to do so, various projects have been taken up under TDIL initiative to provide technological advancements in one's local language.

The objective of the project is to develop various kinds of monolingual and bilingual text to speech synthesis systems for Indian languages, Indian English, to integrate the voices with screen readers, and to make the resources developed as a part of the project. Which in return is available to all groups of people working for corpus generation and research activities through website.

TDIL Projects

Project details as provided by the project teams:

Project 1

Table 3. 6: Project details TTS (Text to speech systems phase ii)

S no	Project Components	Details
1.	Title of Project/Technology: Start Date & Completion Date:	DEVELOPMENT OF TEXT TO SPEECH SYSTEMS IN INDIAN LANGUAGES PHASE II 24/01/2012 – 30/9/2017
2.	Implementing Agencies / Participating Institutes:	IIT MADRAS, IIT MANDI, IIT GUWAHATI, IIITHYDERABAD, IISC BANGALORE, C-DACKOLKATA, C-DAC MUMBAI, C-DAC PUNE, C-DAC TRIVANDRUM, DAIICT GANDHINAGAR, PESIT BANGALORE, SSNCE CHENNAI, IIT KHARAGPUR
3.	Chief Investigator:	Dr Hema A Murthy, professor, IIT Madras

4.	Project Details: Title: Development of Text to Speech Systems in Indian Languages Phase II Project Number: CSE1112129DITXHEMA Project Type: Sponsored Research Project Duration: 24/01/2012 – 30/9/2017
5.	Deliverables/outcome achieved in physical terms: 1. 10 hour speech database with text transcription for one male and one female speaker for 13 Indian languages and the corresponding accented Indian English: Assamese, Bengali, Bodo, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Odia, Rajasthani, Tamil, and Telugu. 2. Unit selection synthesis (USS) voices for different Indian languages. 3. Hidden Markov model (HMM) based speech synthesis (HTS) voices for different Indian languages. 4. HTS voices with STRAIGHT vocoder for different Indian languages. 5. Festival lite (Flite) voices for different Indian languages. 6. Indian English voices for some Indian languages. 7. Bilingual TTS systems in USS and HTS framework. 8. Merlin (a neural network based TTS) based voice for Hindi 9. Integration of USS and HTS voices with screen readers for Windows (NVDA) and Linux (ORCA) platform. 10. A hybrid HMM and group delay (GD) based algorithm for automatic speech segmentation of Indian languages. 11. An automatic speech segmentation algorithm based on deep neural networks (DNNs). 12. A unified Lex and Yacc parser for all Indian languages. 13. A common label set (CLS) for 13 Indian languages. 14. A common question set (CQS) for Indian languages for tree-based clustering in HTS. 15. Four Android applications were developed to make the TTS services available in Android platform: Hindi TTS app, Tamil TTS app, Telugu TTS app, Indic TTS app 16. Tamil learning android application. 17. IndicTTS is a website that hosts all the TTS-related software, database and

	applications under one roof. 18. Android applications with HTS voice. 19. TTS Builder Toolkit (TBT): A High Performance Framework to build TTS

Project 2

Table 3. 7: Project details TTS (Integration of 13 Indian language)

S No	Project Components	Details
1.	Title of Project/Technology: Start Date & Completion Date:	INTEGRATION OF 13 INDIAN LANGUAGE TTS SYSTEMS WITH SCREENREADERS FOR WINDOWS, LINUX AND ANDROID PLATFORMS. 27/03/2018 – 26/06/2019
2.	Implementing Agencies / Participating Institutes:	INDIAN INSTITUTE OF TECHNOLOGY MADRAS, C-DAC MUMBAI
3.	Chief Investigator:	Dr Hema A Murthy, Professor, IIT Madras
4.	Project Details: Title: Integration of 13 Indian language TTS systems with Screen readers for Windows, Linux and Android platforms. Project Number: CSE1819160DEITHEMA Project Type: Sponsored Research Project Duration: 27/03/2018 – 26/06/2019	

5.	Deliverables/outcome achieved in physical terms: 1. HTS voices with speech data for all 13 Indian languages (male and female speakers). 2. 10 hours speech corpus time-aligned at syllable and phone level for 13 Indian language and Indian English datasets (both male and female). 3. Speech data from additional two male and female speakers for the languages: Hindi, Tamil, Gujarati, Malayalam, and Marathi. 4. Monolingual and bilingual TTS synthesisers for all Indian languages and Indian English integrated with ORCA screen readers. 5. Talkback feature in Android platform that support multiple Indian Languages. 6. IndicTTS website with all supporting software, database, applications, and voice models with proper documentation.

3.3.8 DEVELOPMENT OF CORPORA OF TEXTS IN MACHINE READABLE FORM (TEXT CORPUS)

Overview

The text corpora usually represent a large collection of representative samples which are obtained from texts covering different grammatical varieties of languages used.

The main objective of the project was to build annotated parallel corpora in the domain focusing on tourism and health, where Hindi is the source language. The encoding and annotation for the same are as per the global standards.

This research talks about Text Corpus collection of various Languages which initially involved 12 languages (Hindi, Bangla, Punjabi, Oriya, Marathi, Gujarati, Konkani, Urdu, Tamil, Telugu, Malayalam and English) and the features of Text Corpus emphasising upon natural language efficiency for the ease of understanding and collecting information. Various types of Text Corpus are also discussed briefly here.

Introduction

Information today is available in digital world, however, it is accessible to a few who can read and understand a particular language. Language technologies can provide solutions in order to reach out to masses and facilitate the exchange of information across different people speaking different languages. In a country like India, where multi-lingual societies are prevalent, there is a need to standardise the means of communication. Thus, large vocabulary systems have been started to develop as an initiative by the Indian government.

Corpus in its literal sense can be understood as a large collection of writings of a specific kind or recorded remarks. Text corpus thus, is a collection in a written format in different languages. It is directly used in various areas of linguistics such as in the study of syntax, semantics, language teaching, etc. As a knowledge resource, corpus is used in building multilingual libraries, bilingual dictionaries and multilingual lexical resources as a building block to develop various National

Language Processing technologies like that of speech database for Automatic Speech Recognition system, Machine Translation system, OCR, voice recognition system, and text to speech. Hence, evaluation of corpora is essential and a key factor.

Text corpus is a large and structured set of texts which in today's time is usually electronically stored and processed. In corpus, they are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory.

Text Corpus Collection

Indian Languages Corpora Initiative started by Technology Development for Indian Languages (TDIL) programme of Ministry of Communication and Information Technology for building parallel corpora for major Indian languages including English. The main objective of the project was to build annotated parallel corpora in the domain focusing on tourism and health, where Hindi is the source of language. The encoding and annotation for the same are as per the global standards.

As mentioned above, the text corpora usually represent a large collection of representative samples which are obtained from texts covering different grammatical varieties of languages used. To simplify, corpora could be defined as **C**apable **o**f **R**epresenting **P**otentially **U**nlimited **S**election of Texts.

When we talk about corpus, the same can be written or spoken text, new or old texts, and monolingual, bilingual or multilingual text. Further, it can be obtained from books, newspaper, journals, and speeches. There are various types of corpora such as written corpora, speech corpora, image corpora, parallel corpora, reference corpora, monolingual corpora, bilingual corpora, untagged corpora, tagged corpora, etc. Evaluation strategy differs as per the different type of corpora.

Features of Text Corpus

The main purpose of a corpus is to verify a hypothesis about language - for example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus linguistics deals with the principles and practice of using corpora in language study. Some of the key features of an efficient text corpus shall include that of:

- It should be able to represent natural languages efficiently for the ease of understanding and as a collection of information.
- Keeping in the mind the vastness and the application of the text corpus it is thus necessary for it to be largely available and in a balanced format for all the disciplines.
- The objective should be, and is, to capture almost all the linguistic features of the languages.
- Ease of availability and adaptability should be empowered, wherein the information could be retrieved easily and used by the end users.
- The technology should be independent and subjected to many kinds of empirical investigation and analysis, and
- As a resource, it should be available and good enough for developing sophisticated language processing tools.

For the development of phonetically rich sentences, one of the important decisions which needs to be made is the choice of huge text corpus source from which sub-text shall be extracted. Thus, the reliability and coverage are the major factor of text corpus which in turn lays foundation for various other language processing tools. The corpus, then, should be unbiased and large enough to convey the entire syntactic behaviour of the language.

Types of Text Corpus

Annotated Text Corpus

An annotated corpus may be considered to be a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete annotation.

Comparable (reference) corpus

A comparable corpus is a set of two or more monolingual corpora whose texts relate to the same topic, however, they are not translations of each other, and therefore, there are not aligned. When users search these corpora, they can use the fact, that the corpora also have the same metadata. An example of comparable corpora in Sketch Engine is CHILDES corpora or various corpora made from Wikipedia.

Monitor Corpus

A type of corpus which is a growing, non-finite collection of texts, of primary use in lexicography. Monitor corpus reflects language changes in a constant growth rate of corpora, leaving untouched the relative weight of its components (i.e., balance) as defined by the parameters.

Monolingual Corpus

Monolingual corpus is the most frequent type of corpus. It contains texts in one language only. The corpus is usually tagged for parts of speech and is used by a wide range of users for various tasks from highly practical ones, e.g., checking the correct usage of a word or looking up the most natural word combinations, to scientific use, e.g., identifying frequent patterns or new trends in language. Sketch Engine contains hundreds of monolingual corpora in dozens of languages.

Parallel (aligned) Corpus

A parallel corpus consists of two monolingual corpora. One corpus is the translation of the other. The user can then search for all examples of a word or phrase in one language and the results will be displayed together with the corresponding sentences in the other language. The user can then observe how the search word or phrase is translated.

Reference Corpus

A type of corpus that is composed on the basis of relevant parameters and should include spoken and written, formal and informal language representing various social and situational strata.

Spoken Corpus

A type of corpora that contain texts of spoken language.

Unannotated Corpus

A type of corpora that are in raw states of plain text; opposed to annotated corpora.

Speech Corpus

A large collection of audio recordings of spoken language. Most speech corpora also have additional text files containing transcriptions of the words spoken and the time each word occurred in the recording.

Speech Corpora can be divided in two types:

- Read Speech: which includes Book excerpts, Broadcast news, Lists of words, Sequences of numbers.
- Spontaneous Speech: which includes Dialogs & Meetings - between two or more people; Narratives - a person telling a story; Map-tasks - one person explains a route on a map to another; Appointment-tasks - two people try to find a common meeting time based on individual schedules.

TDIL Projects

Project details as provided by the project teams

Table 3. 8: 7 Project details Text Corpora

S No	Project Components	Details
1.	Title of Project/Technology: Start Date & Completion Date:	INDIAN LANGUAGES SPEECH RESOURCES DEVELOPMENT FOR SPEECH APPLICATIONS Start Date - 21.12.2016 Completion Date – 20.06.2019
2.	Implementing Agencies / Participating Institutes:	CDAC NOIDA, CDAC KOLKATA, KIIT GURGAON
3.	Chief Investigator:	a. Sunita Arora, Joint Director b. Department: Speech and Natural Language Processing Lab (SNLP) CDAC Noida
4.	**Project Details:** This is a consortium-based project and the members are C-DAC Noida, C-DAC Kolkata and KIIT Gurgaon, headed by C-DAC Noida. The aim of the project was to develop Speech Database for Indian Languages, viz., Hindi, Bengali and Indian English. The database has been developed and contains audio files and text transcriptions of 1500	

	speakers for Travel, Agriculture and General Domain, for different recording environments. The recording environments are studio, home/office and roadside to capture the impact of different types of noise on speech. The text corpus is phonetically balanced and phonetically rich. This database can lead to successful developments of speech technologies like Automatic Speech Recognition Systems and Text to Speech Systems for Indian Languages.
5.	Deliverables/outcome achieved in physical terms: 1. Speech corpus for ASR of Indian languages (Hindi, Bangla, Indian English spoken by native Hindi speakers) 2. Speech corpus for Speech Synthesis of Indian languages (Hindi, Bangla) 3. Pronunciation Lexicon 4. Meta Data information of speakers 5. Corpus Documentation 6. Language Specifications (LSP)

Indo Wordnet

Introduction

The multilingual nature of India necessitated the creation of Wordnets and lexicons of high quality and coverage to further facilitate communication between people from different linguistic groups. The Indo Wordnet is a lexical database for various major Indian languages such as Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu, and Urdu, which was created applying the expansion approach from Hindi and English Wordnet. The project was a culmination of the Dravidian WordNet, North-East WordNet and Indradhanush projects, all of which were funded by the TDIL programme.

This was made possible by the collective endeavours of a consortium of members including Goa University; Indian Institute of Technology Bombay, Mumbai; Indian Statistical Institute, Kolkata; Dharmsinh Desai University, Nadiad; University of Kashmir, Srinagar; University of Hyderabad, Hyderabad; Punjabi University, Patiala; Thapar University, Patiala; and Jawaharlal Nehru University, New Delhi.

Benefits

The Indo WordNet today is not a compilation of lexical items in some order but it is rather a repository where various lexical items are systematically categorised based on certain predetermined aspects, such as part-of-speech and lexical type. It is a useful free lexical resource available to the masses in general and is the first of its kind in Indian languages. It is an essential part for facilitating the creation of machine translation system for Indian languages, as well as English, and in enabling query translation for systems like Cross Lingual Information Retrieval (CLIR) systems. The bilingual and multilingual dictionaries that are in the state of compilation from the WordNet are useful for manual as well as statistical machine translation.

The Indo Wordnet has also made it easier for people to learn a new Indian language. This project has also made it possible to create an extraordinary alliance between computer

scientists and linguists. This interdisciplinary approach has facilitated understanding across both groups. It has helped the scientists comprehend the nature and complexities of computation and at the same time helped computer scientists recognise the intricacies of the way natural language operate. The Indo WordNet has thus emerged as a digital lexical knowledge base for 18 Indian languages.

3.4 SUMMARY

TDIL has recognised a number of language technologies as being important for the development of ICT in India. Through its assorted institutions, it has conducted extensive research in optical character recognition, automatic speech recognition, online handwriting recognition, machine translation, cross lingual information access, speech synthesis, and text corpus organisation. Its primary aim has always been to consolidate digital information and make it available to the public at minimal cost. This chapter has introduced the aforementioned technologies and then listed out the assorted projects undertaken by TDIL departments. It is clear that the Indian government is taking steps to promote a culture of technological awareness and hence usher the country into a promising digital future.

Chapter 4: Understanding TDIL Standardisation

4.1 OVERVIEW

In order to develop a nation-wide programme such as TDIL which aims to bring people from various backgrounds together, there is a need and necessity to create a standard form. In a business forum, standardisation is a framework of agreements to which all relevant parties in an industry or organisation must adhere to, to ensure that all processes associated with the creation of a good or performance of a service are performed within set guidelines. However, in this chapter, we try and understand the need and working of standardisation on a nation-wide platform. We discuss and elaborate more about understanding standardisation, the need and benefits of standardisation along with its due processes involved in it.

4.2 INTRODUCTION

Standardisation in its literal terms is the process of implementing and developing technical standards based on the consensus of different parties that include firms, users, interest groups, standards organisations and governments. Standards or a unified platform support our everyday life. We owe much of progress in modern “tech” world to standardisation. Without

standards, our lives would have been a bit difficult in terms of doing basic things. Years ago, society recognised the need of having a unified platform in terms of length, weight, height to unanimously deliver the outcomes. With technological progress, the need for standardisation grows. The rapid progress in the area of information and technology (ICT) could have been achieved only because of the unification of such information. Thus, standardisation and particular simplified standards boosts up the progress and create a basis for the technology to evolve and develop. Though important, ICT standardisation and its methods remain a topic that is not easily accessible. It seems that this field is becoming increasingly limited to the expert and remains mysterious to the nonexpert. Standardisation, in particular in the area of ICT, deserves more attention.

For a vast, multilingual country like India, standardisation is the only means to unite the diversity of the Indian language arena, to develop and maintain best practices in the field of languages. Currently, India has 22 constitutionally recognised languages and 12 scripts, it is a necessity to develop a common platform of standardisation to develop and maintain best practices in the field of languages. The aim of the TDIL Programme, initiated by the Ministry of Electronics and Information Technology (MeitY) is to achieve communication without language barrier in the field of ICT.

4.3 NEED AND BENEFITS OF STANDARDISATION

Today the technology advancements in India is at peak. However, according to stats, most of the benefits of the applications of ICT are not being able to be utilised by the masses. The only gap between the technology and its application is because of the language barrier. Most, or to say all, the ICT development and applications are available in English language. Keeping in mind the language strata of India, only 5-7 percent of the population can speak, read and write in English Which, in turn, not only hinders with the growth of ICT in the country but also with the development of the society. In order to curb it, the TDIL Programme initiative was taken at par. The aim is localising the set the applications of ICT to be made available for the masses for a better development of the society. Various web applications and programmes are being translated into Indian local languages for them to made available for the citizens, wherein the technology is made accessible for them and much more user-friendly.

Developing technologies to use Indian languages in computers was the basic need. The spectrum of development need for any language is a very huge task and when the challenge is to work on language requirements across India, geographies, people and diversities, it becomes even more difficult. There was a dire need to create standardisation for using Indian languages to unite the diversity and bringing the different stakeholders on the same page. Standardisation was required right from the script symbols in Unicode, efficient and best use-based keyboards, fonts, Web and mobile standardisation, W3C standards compliance, SMS, Transliteration and Speech Resource Standards, Web and mobile standardisation for Indian languages. Currently, India has 22 constitutionally recognised languages and 12 scripts. Common standardisation platform was very important to enable the masses with the power of ICT. There is a need to interact and be compliant with the global standards and become part of the global consortiums for the common standards. In turn, common standards will enable the industry and the developer community in India and globally to build applications, products and services for the consumption of Indian masses. Produced anywhere and consumed anywhere, ultimately helping the Indian community.

The TDIL programme initiated by the Ministry of Electronics and Information Technology (MeitY) is an apex program for the development and maintenance of standardisation of Indian languages, helping achieve communication without language barrier in the field of ICT. To maintain the universal sustainability and development of India and its living languages in the ICT system, TDIL is working in the field of standards so that its key pillars of Indian Languages scripts would be preserved while developing application for mass usage. Individually, each and every process has a larger impact on the programme. The necessity is to create the platforms user friendly and thus, require a standard form.

4.4 PROCESSES OF STANDARDISATION IN TDIL PROGRAMME

TDIL is working for the development and enhancement of the standards in the following field with various international organisations. The study further penetrates into understanding each and every tool in depth and its needs and benefits.

4.4.1 UNICODE STANDARDS

Unicode is a computing industry, a non-profit organisation devoted to developing, maintaining, and promoting software internationalisation standards and data. It consists of standards for the

consistent encoding, representation and handling of text expressed in most of the world's writing system. The Unicode Standard consists of a set of code charts, an encoding method and set of standard character encodings, a set of reference data files and a number of related items including of character properties like rules for normalisation, decomposition, collation, rendering and bidirectional display order. The Unicode Standard specifies the representation of text in all modern software products and standards and actively develops standards in the area of internationalisation including defining the behaviour and relationships between Unicode characters.

The consortium works closely with W3C and ISO to maintain International Standard synchronised with the Unicode Standard. MeitY is a voting member of the Unicode Consortium. All twelve Indian scripts, namely, Bengali, Devanagari, Gujarati, Gurumukhi, Kannada, Malayalam, Meethi-Mayek, Odia, Ol-Chiki, Perso-Arabic, Tamil and Telugu are represented in Unicode.

UNICODE TYPING TOOL

This is a typing software, which enables typing of Indian Languages in editors of Windows based applications with Unicode compliant font. It supports typing in Assamese, Bangla, Boro, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Marathi, Manipuri, Nepali, Odia, Punjabi, Sanskrit, Sindhi, Santali, Tamil, Telugu, and Urdu.
Along with Sakal Bharati Font, this typing tool contains two open type fonts for each language and the list of fonts is available in supporting documents. On-screen keyboards for each language are also provided in this tool to make typing easier.
Unicode Typing Tool now supports iWriting - Predictive typing feature with Inscript Keyboard which currently supports 10 languages like Assamese, Bangla, Boro, Hindi, Marathi, Odia, Punjabi, Tamil, Telugu, and Urdu. It provides multiple options for auto-completion of word. It also comes with intelligent self-learning feature.

There are two projects of Unicode which are evolving under the umbrella of TDIL:

Character Encoding

Character encoding is the numeric value assigned by Unicode for all the characters and symbols used in writing scripts using UTF-8 encoding. Character encoding standards not only identity each character and its numeric value, or code point, but also how this value is represented in bits.

The four level of the Unicode Character Encoding Model can be summarised as:

- Abstract Character Repertoire (ACR): the set of characters to be encoded, for example, some alphabet or symbol set.
- Coded Character Set (CCS): a mapping from an abstract character repertoire to a set of non-negative integers.
- Character Encoding Form (CEF): a mapping from a set of non-negative integers that are elements of a CCS, to a set of sequences of particular code units of some specified width, such as 32-bit integers.
- Character Encoding Scheme (CES): a reversible transformation from a set of sequences of code units (from one or more CEFs to a serialized sequence of bytes)

Unicode Standard follows a set of fundamental principles for Character Encoding: Universal repertoire, Logical order, Efficiency, Unification, Characters, not glyphs, Dynamic composition, Semantics, Stability, Plain Text and Convertibility.

Common Locale Data Repository

The Common Locale Data Repository (CLDR) provides key building block for software to support all the languages of the world and is maintain and modified by UNICODE consortium. CLDR is the largest and most extensive standard repository of locale data. Its goal is to gather basic linguistic information for various "locales," essentially combinations of a language and a location, so that this data will be used for software internationalisation and localisation. This is achieved by adapting a software to the conventions of different languages for such common software tasks as formatting of dates, times, time zones, numbers, and currency values; sorting text; choosing languages or countries by name; and many others. The basic lists that CLDR gathers are dates formats, time zones, number formats, currency formats, measurement system, collation specification: sorting, searching, matching, translation of names for language, territory, script, time zones, currencies, script and characters used by a language.

4.4.2 INDIAN RUPEE SYMBOL

The Indian Rupee, on 15thJuly, 2010, got its own unique and standard symbol with the approval from the Union Cabinet, whose numerical representation was now to be a mix of Devanagri ‘Ra’ and Roman ‘R’. With this new symbol, Indian Rupee became the 5th currency globally to have a symbol after four other currencies – Japanese Yen, US Dollar, British Pound and Euro. Within three months of its selection, this symbol, with the efforts of Government of India, Unicode Consortium and ISO, was incorporated as the UNICODE standard and was available at code

point U+20B9 in the currency symbols block of Unicode version 6.0. There was a consensus among the various stakeholders to use the combination of 'AltGr' and the numeral '4' as the typing sign with regards to placement of the symbol on the keyboard.

4.4.3 SCRIPT BEHAVIOR

Script Behaviour defines the manner in which a script of a given language is to be written and, especially in the case of Indian languages, the ways in which ligatures have to be constructed. The script grammar once defined and mandated by the authoritative body of the state creates a standard which will serve as a guide for developers of fonts. Eventually it will also determine the manner in which all printed as well as digital material is prepared, thereby slowly bringing in one standardised manner of representing the given language.

4.4.4 LANGUAGE RESOURCE DEVELOPMENT

When we talk about language development and processing, the first step that needs to be taken and adhered to is the development of the resources. The need is to create and develop a data set which contains information in regard to various languages. Following the same, the TDIL programme of MeitY has aimed to develop a language resource which contains information such as corpus, scripts, etc., of various Indian languages.

The TDIL programme took the initiative to create various resources for Indian languages which are as follows:

Ontology

Ontology refers to representation or formal naming and definition of properties, categories and relations between concepts and data of one, many or all domains. It is an explicit formal specification of how to represent the objects, concepts and other entities which exist in some area of interest and the relationship that holds them. It also includes hierarchical structuring of knowledge about things by sub categorising them according to their essential qualities.

Corpora

Corpora are the main knowledge base in corpus linguistics. It is essentially a large and structured set of texts which are used to do statistical analysis and hypothesis testing, checking and validating linguistic rules. These are part of computational linguistics. Speech recognition and machine translation needs analysis and processing of various types of corpora, to create part of speech tagging and morphs, semantics, etc. Corpora have further structured levels of analysis; such corpora are usually called Treebanks or Parsed Corpora. Corpora is further divided as:

- **Text Corpora**

A collection of writings used for linguistic analysis.

- **Speech Corpora**

A collection of recorded remarks used for linguistic analysis. Speech Corpus for Assamese, Bangla, Hindi, Manipuri, Marathi and Punjabi are available for Indian Researchers.

Lexical Resources

Based on their nature and function, lexical resources can be further divided as:

- **Pronunciation Dictionary**

A reference book containing an alphabetical list of words, with information given for each word, usually including meaning, pronunciation and etymology. There is no specified standard for creating Dictionary structure. XML is now mostly used and recommend way of creating structure as it is more logical and useful for creating web-based dictionaries.

- **Thesaurus**

A Thesaurus is a book of selected words or concepts, such a specialised vocabulary items of a particular field. It often contains synonyms, and other semantically related words including related and contrasting words and antonyms.

- **Term Bank**

A stock of terms used in a particular profession, subject or style.

Linguistic Analysis

Linguistics is the study of the nature, structure, and variation of language or words, and the words are analysed on the basis of phonetics, phonology, morphology, syntax, semantics, sociolinguistics, pragmatics and discourse analysis:

- Phonetic Analysis- The sounds of speech production, combination and representation by written symbols.
- Phonological Analysis- Study of speech sounds of a language with reference to distribution and patterning.
- Morphological Analysis- Deals with root / base form of the word and the morphemes affixed to it.
- Syntactic Analysis- It deals with grammatical analysis of sentences or discourse structure.
- Semantic Analysis- Concept-based analysis.
- Sociolinguistics- it deals with the study of the sociological aspects of language
- Pragmatics- It is concerned with the use of language in social contexts and the ways people produce and comprehend meanings through language
- Discourse Analysis- Analysis of the discourse structure by using knowledge of the world.

Formalism

Refers to the syntax of language or well-formed formulas of grammar, such that the inference rules can be derived for language processing. Principally a study of theoretical framework or syntax of language for computational linguistic analysis.

4.5 STANDARDISATION BODIES

A Standard's organisation or a standardisation body is an organisation that aims at developing, coordinating, propagating, revising, altering, reissuing, interpreting and producing technical standards. A standard works to provide specifications for various products, services and systems, so as to assure quality, safety and efficiency. They also play a major role in facilitating international trade. A standards body can be categorised on the basis of their role, position, as well as the extent of their influence on the local, national, regional and global standardisation sphere. The TDIL currently is working in collaboration with various national and international standards organisations such as the International Organisation of Standardisation (ISO), UNICODE, World Wide Web Consortium (W3C), European Language Resources Association (ELRA), as well as the Bureau of Indian Standards (BIS).

4.5.1 INTERNATIONAL ORGANISATION FOR STANDARDISATION (ISO)

The International Organisation for Standardisation or the ISO is the largest, independent, non-governmental international standards organisation. It is an international standard setting body which is made up of members from various national standards organisations with over 164 member nations. It works to form a link between the public and private sectors. The ISO has a central office in Geneva, Switzerland which regulates the working within the organisation. The TDIL, under the Department of Electronics and Information Technology (now MeitY), represents India at ISO as an issue based representative body.

4.5.2 UNICODE

The UNICODE standard is a modern universal coding standard that represents the way individual characters are depicted in text files, web pages, and other forms of documents. It works to illustrate letters and symbols that are frequently used in present day digital as well as print media. It allows for the data to be moved from many different platforms, devices and applications without corruption. Unicode today has become the highest standard for classifying characters in text in nearly any language. MeitY is a member of the Unicode Consortium with voting rights.

4.5.3 WORLD WIDE WEB CONSORTIUM (W3C)

The World Wide Web Consortium is the main international standards organisation for the World Wide Web that works with the motive of improving the web and leading it to its full potential. It is made up of member organisations which sustain a full-time staff with the aim of working together in the advancement of standards for the World Wide Web. The consortium also works on developing software and serves as an open forum for discussion about the web. India is a full voting member of its current 444 members.

4.5.4 EUROPEAN LANGUAGE RESOURCES ASSOCIATION

The European Language Resources Association (ELRA) is a non-profit organisation which was founded in 1995 with the mission to make Language Resources (LRs) for the Human Language

Technologies (HLT) available to the community at large. Their mission, since its inception, has enlarged and now also include production and identification of new language resources as well as evaluation of language engineering tools. The TDIL is voting member of the ELRA on behalf of the Government of India.

4.5.5 BUREAU OF INDIAN STANDARDS (BIS)

The Bureau of Indian Standards (BIS), formerly known as the Indian Standards Institution, is the national Standards body of India that works under the Ministry of Consumer Affairs, Food and Public Distribution, Government of India. It was established by the Bureau of Indian Standard Act, 1986, that came into effect on 23rd December 1986. The organisation works to formulate, recognise, and promote Indian standards. It's headquartered in New Delhi, with 5 Regional offices and several Branch offices throughout the country. The organisation was also a founder member of the International Organisation for Standardisation (ISO).

4.6 SPEECH RESOURCES STANDARDS

Speech Processing is the study of speech signals and the methods used for processing those signals. It evaluates human speech using digital signal processing techniques. The facets of speech processing may vary from acquisition, manipulation, storage, transfer and output of speech signals depending upon the focus of analysis, where the input is called speech recognition and the output speech synthesis.

Speech processing provides dynamic tools for improving the interaction between humans and machines, and between humans using machines. It can further be built upon using the Natural Language Processing (NLP) technology which uses a computer program or software to understand and manipulate the natural human language. It is a component of Artificial Intelligence (AI) which aims to achieve human ability and capacity to understand and process the content of the human language, and to facilitate the translation of a spoken sentence from one language to another, as well as various other smart linguistic applications. Speech tools created using the NLP are also a good medium for yielding information access interface to people with various visual and cerebral disabilities, as it uses human language to extract meaning and make decisions accordingly, based on the information provided. They are also useful in the present age of digitisation where the NLP technology is present and used in various machine translation applications, that allows us to overcome barriers of communicating with people who speak different languages, which is essential for a multilingual country like India.

Speech resources such as these are, therefore, a detrimental component for the development of speech-based systems in India. The advancement of these resources also requires various standards and methodologies which vary according to the different aspects of speech processing that is being put into use. The TDIL under the Ministry of Electronics and Information Technology, Government of India, has, thus, taken various initiatives towards the development of speech resources for Indian Languages.

4.7 SMS STANDARDS

Short Messaging Service (SMS) has been a remarkable technological tool used for communicating in the modern day. Information is easily disseminated and communicated via this platform to a large number of people, nationally and globally. Since a large number of people use this medium for communicating, it is crucial for it to inculcate standard rules in its delivery. The technology has seen an accelerated growth in rural India and since English literacy is still low, it becomes crucial for an efficient Indian languages support mechanism.

Multilingual data handling is vital through different layers in Mobile Technology. The encoding scheme for data transmission should consider the following parameters:

- The data encoding scheme should support all possible characters, character combinations in Indian Languages as per Unicode standard.
- There should be a provision to change languages within a single message.
- The encoding should be flexible for future Unicode standard (TDIL).

For the benefit of approximately 750 million mobiles across India, the TDIL programme of MeitY is working on developing standards in the Indian Languages SMS. Service providers can use this tool to increase their user base and transmit important information conveniently.

Presently, three prevalent SMS encoding schemes are available in India:

- Indian Script Code for Information Interchange (ISCII) based 7-Bit encoding
- 7-bit default alphabets as per GSM standard
- UTF-8

4.8 TRANSLITERATION STANDARDS

As majority of population know more than one language, they understand the spoken or verbal communication, however when it comes to scripts or written communication, the number diminishes. Thus, a need for transliteration tools which can convert text written in one language script to another script arises. Transliteration should not be confused with translation, which involves a change in language while preserving meaning. Transliteration is mapping of pronunciation and articulation of words written in one script into another script.

A transliteration method also requires knowledge to have the correct pronunciation. Thus, transliteration is meant to preserve the sounds of the syllables in words. Transliteration is helpful in situations where one does not know the script of a language but knows to speak and understand the language nevertheless.

Guidelines in Unicode for transliteration standards

1. Complete: Every well-formed sequence of characters in the source script should transliterate to a sequence of characters from the target script, and vice versa.

2. Predictable: The letters themselves (without any knowledge of the languages written in that script) should be sufficient for the transliteration, based on a relatively small number of rules. This allows the transliteration to be performed mechanically.
3. Pronounceable: The resulting characters have reasonable pronunciations in the target script. Transliteration is not as useful if the process simply maps the characters without any regard to their pronunciation.
4. Reversible: It is possible to recover the text in the source script from the transliteration in the target script. That is, someone that knows the transliteration rules would be able to recover the precise spelling of the original source text.

4.1 KEYBOARD STANDARDS

Indian Language Keyboard is categorised into three parts- Inscript (Indian keyboard), phonetic English keyboard, typewriter keyboard. The characters of Indian language alphabets are divided into Consonants, Vowels, Nasals and Conjuncts. Every consonant represents a combination of a particular sound and a vowel. The vowels are representations of pure sounds. The Nasals are characters representing nasal sounds along with vowels. The conjuncts are combinations of two or more characters. The Indian language alphabet table is divided into Vowels (Swar) and Consonants (Vyanjan).

Inscript (Indian Keyboard)

The INSCRIPT (Indian Script) Keyboard Layout was standardised by the Department of Electronics (DOE) in 1986 with a subsequent revision in 1988. The INSCRIPT keyboard layout was declared as a National Standard by Bureau of Indian Standard (BIS) in 1991.

This keyboard overlay is phonetic in nature and has a common layout for all the scripts provided with this software. It contains the characters required for all the Indian scripts, as defined by the ISCII character set. The Indian script alphabet (ISCII) has a logical structure derived from the phonetic properties of the Indian scripts. The INSCRIPT overlay mirrors this logical structure.

With the advent of Unicode, a few new characters were added to each code-page, characters for which the INSCRIPT keyboard layout standard had not made any provision. In addition, the concept of ZWJ and ZWNJ, as well as that of normalisation were also added. These new features had a marked repercussion on storage as well as input, and an urgent need was felt for a revision, whereby each and new character introduced in Unicode would be accommodated on the keyboard and a uniform manner of entering data as well as storing data would be devised.

This layout uses the standard QWERTY 101 keyboard. The mapping of the characters is such that it remains common for all the Indian languages (written left to right). This is because of the fact

that the basic character set of the Indian languages is common. In the INSCRIPT keyboard layout, all the vowels are placed on the left side of the keyboard layout and the consonants, on the right side. The placement is such that the characters of one varg are split over two keys. The splitting of the word into keystrokes is based on the phonetic spelling of the word. The sequence required for typing a word is same as the sequence in which the characters of the word are pronounced.

Phonetic English Keyboard

Phonetic English keyboard overlay has the Indian script alphabets phonetically assigned to that of English alphabets on the IBM-PC QWERTY overlay. This keyboard is useful for people who are acquainted with the language but know how to use computer using English.

The Phonetic keyboards are useful for those who can speak but cannot write in their mother-tongue and for those who are comfortable with the QWERTY keyboard and do not want to key in a text using the INSCRIPT keyboard.

Typewriter Keyboard

Typewriter Keyboard overlay functions in the same manner as the manual typewriter. Users accustomed to working with a typewriter can use this facility with minimal learning and training time.

This type of keyboard has minimal set of aksharas consisting of the basic vowels and consonants together with the matras so that text can be prepared conforming to the writing system for the language. The location of the keys for the vowels and consonants on a regional language typewriter is specific to the language and the data entry method would be different for different languages.

4.2 W3C

Web technology is an integral part of human society. The World Wide Web is a system of interlinked hypertext documents accessed via the internet. The aim is to make Web as engaging as possible and expanding its horizons to enhance communication, governance ensuring global transparency, and participation and collaboration into key communities such as industry, political, social and other communities. To lead the Web to its full potential as a forum for information, commerce, communication, and collective understanding, an international community called the World Wide Web Consortium (W3C) was created to develop interoperable

technologies (specifications, guidelines, software, and tools). The World Wide Web Consortium (W3C) is the main international standards organisation for the World Wide Web (abbreviated WWW or W3). The W3C member organisations work together in the development of standards for the World Wide Web.

W3C India Office is a full voting member of W3C and is Advisory Committee Representative. It proposes to work in close collaboration with all stakeholders of academia, govt., industry and industry associations.

The futuristic and the very long-term goals of W3C India office are to enable all W3C recommendations with 22 Indian languages so that seamless web for every Indian can be achieved.

1. W3C India Office will be apex body of W3C activities in India and would act as Single Window for Bi-directional communication between Stakeholders and W3C Consortium.
2. Education and Outreach to all stakeholders, Promotion and proliferation of W3C Standards and communicating national level feedback to W3C for the existing and future standards of W3C.
3. National Level Special Interest Groups (SIGs) in Technology Intensive areas will be setup in some priority areas to evolve National Recommendations.
4. W3C India Office will run W3C India Portal and also circulate W3C News-letters to all Stakeholders in India. Bilingual News-Letters will be brought out to reach new segments of Stakeholders from International perspective.
5. W3C India office is presently looking after Internationalisation, Voice browse, Styling, Web Accessibility and Mobile Web as the standards with respect to Indian languages.

4.3 WEB STANDARDISATION INITIATIVE

India is very rich and wide in regard to linguistic diversity. India has 22 constitutionally recognised languages and 12 scripts. In India, there is one language might have originated from many scripts and many languages might have originated from one script and they are culturally different depending on the region, though using the same script for different languages. Even there is wide difference for the same language across different country. There are also major issues related to Indian Languages like complex orthographic representation, non-uniform mapping between script and languages, the formation of complex glyph sets and other NLP problems. To accelerate the percolation ICT in Indian languages, it is necessary to enable web

technology in all 22 Indian languages. This requires addressing internationalisation challenges with respect to present and future web standards.

The vision of WSI is to enable all Web related standards with 22 Indian languages so that we can achieve seamless Web for every Indian. The objective of this initiative is to promote and adopt various internationalisation web and internet recommendations/ best practices among developers, application builders, and standards setters, and to encourage inclusion of stakeholder organisations in the creation of future recommendations.

WSI is currently focusing on the development of Internationalisation requirements with respect to all 22 Indian languages in the following areas:

1. CSS & Digital publishing
2. Mobile web
3. Semantic Web
4. Speech technology
5. WOFF (Web Open Font Format)
6. Internationalisation Tag Set

Under WSI, W3C Indic Text layout task force has been constituted and is working directly under the W3C Internationalisation Working Group. The Chairperson of this Task force is Ms. Swaranlata. The aim of this task force is to collect information about specific use cases in India for technologies defined in W3C specifications for Indian languages, and to report the results of its activities as a group back to the Internationalisation Core Working Group, as well as to other relevant groups and to the W3C members and the community.

WSI is produced W3C First Public Working Draft of Indic Layout Requirements on behalf of the Indic Layout Task Force, part of the W3C Internationalisation Interest Group and Digital publishing Interest Group.

4.4 CONCLUSION

We have already discussed various processes and the need of such form of standards. Standardisation delivers measurable benefits and establishes an international consensus on terminology, thus making technology transfer easier and safer. Various processes of standardisation include that of Unicode Standards, Standard form of Indian Rupee Symbol, Script Behaviour, Language Resource Development, role of various Standardisation Bodies,

Speech Resource development and its standards, SMS, Transliteration, Keyboard, role of W3C in TDIL and Web Standardisation Initiative. Each and every process is unique in its own functioning and structure, thus has a greater impact in the development of TDIL programme.

4.5 SUMMARY

This chapter looked at the process by which different technical standards were adopted in order to make TDIL's research data universally comprehensible and accessible. Different standardisation bodies were consulted and considered, including the:

- International Organisation for Standardisation
- Unicode Consortium
- World Wide Web Consortium
- European Language Resources Association
- Bureau of Indian Standards.

Standards were set to define how text should be encoded, how text and speech corpora should be organised, how SMS should be formatted, how Indian web pages should be represented on the internet, and so on. In making its programmes conform to global conventions on design and execution, TDIL has taken a major step towards not only globalizing its content, but also gaining international recognition.

Chapter 5: Understanding TDIL-DC Portal Deployment and Usage

5.1 INTRODUCTION

The TDIL portal, in accordance with its primary aim, hosts various linguistic resources, tools and applications provided by the consortium. These applications covers a wide spectrum of research areas in Language Technology, viz., TTS(Text to Speech), SPTIL Morph Analyser, PLS(Pronunciation Lexicon for Indian Languages), OHWR(Online Hand Writing Recognition), OCR(Optical Character Recognition), MT(Machine Translation), and ASR(Automated Speech Recognition). The website also acts as a discussion forum and a research community builder that provides access to research papers by various researchers, professors, developers of numerous organisations and academic institutes who have worked and assessed the development of various initiatives undertaken by TDIL. They deal with the various ways in which the programme can approach the problems, challenges and limitations in language technology. The portal also gives access to some research papers and research journals: Vishwabharat, Language in India, and Language Technology, which delve critically into the existing achievements, the developing technology as well as the potentiality of language technology and its role in empowering the people to overcome the linguistic barriers they face on the way of their access to technology.

Figure 5. 1: An overview of the TDIL Data centre

The TDIL data centre on www.tdil-dc.in, serves as the main depository for the various Indian Language tools, technologies, standards as well as the different funded projects that are being backed under the programme by the Ministry of Electronics and Information Technology (MeitY). These include the various linguistic tools and resources such the text-speech-image corpora, dictionaries, applications, websites, software, mobile apps, fonts, etc. on the portal. The Data Centre was established with the intention to take language computing to the next level by providing a unifying platform for decision makers, users, domain researchers, linguists as well as computational experts. The portal thus provides basic information about all the tools and software in a simple yet lucid manner about the different functions of the applications and e-portals along with technologies being used and worked upon for the same purpose.

For a bird's eye view of the network security system in place at a typical data centre, refer to Fig. 5.1. The portal also offers various guidelines and FAQs regarding the practices for Indian Language computing for official Indian languages as well as language support on different platforms and applications.

5.2 HISTORY AND DEVELOPMENT

The TDIL portal on www.tdil.dc.in was launched in November 2010 with the aim of bringing about a community platform for Researchers, Developers and end users of Indian technology where they could easily share and access the growth and resources in the Indian Language Technology area. The data centre was the culmination of a large body of efforts initiated under the TDIL programme since the inception of this project. The portal worked to bring together these efforts and developments to the people at large to make them aware of the various language tools, technologies, and applications like the text-speech-image corpora, dictionaries,

Machine translation system, OCR, Cross Lingual Information access, Text to speech, Sanskrit tools and various other services.

This portal for Indian Language and Technology Proliferation Deployment Centre (ILTPDC) has, since its inception in 2010, been hosted by the Centre for Development of Advanced Computing(C-DAC) Pune. It has, from the time of its deployment, undergone two phases, Phase I and Phase II. Both of these have worked to facilitate the dispersion of Indian Language computing and availability of IT resources in Indian languages which in turn assures that the various e-governance applications reach the masses.

Phase I

This phase began with the launch of the website in November 2010. It worked to showcase the various outcomes of Language Technology in India and for Indian Languages till that time and providing a space for people from different linguistic backgrounds to be able to access it. This phase, which spanned over a period of 3 years, endeavoured to make regular updates to the data centre following the various developments in the field and to continue bringing Indian language tools and technologies under one roof. (It also created a heterogeneous environment where various heterogeneous applications, contents are offered as web services.) These efforts made it possible for the TDIL data centre to become a successful community portal for researchers, developers and end users. This sharing of knowledge has also contributed to further research and development into the Indian Language technology area.

Phase II

This phase for the Indian Language Proliferation and Deployment Centre began in December, 2014, and is expected to span over the time of 5 years. This phase aims to redesign the portal wherever necessary in addition to the scaling up and management of the existing data centre created and maintained by C-DAC, Pune. It also plans to host the TDIL data centre at NIC, Pune and creating a Disaster Recovery (DR) site at NIC, Delhi. It is also its agenda to host the "Localisation Project Management" to aid the translation of WIKI contents and to design, upgrade and manage the TDIL website in all 22 constitutionally recognised languages. Phase II also proposes to create a mode of interfacing with various book publishers, educational institutes, newspaper publishers and as many sources of the publishers or written text to ensure that their content is also translated and put on the web in an interoperable manner which will lead to the further proliferation of Indian language content on the web.

5.3 APPLICATION SHOWCASE

5.3.1 SANDHAN (CROSS LINGUAL SEARCH)

It was initiated as a mission mode project under the TDIL programme with the objective of developing a monolingual search system for tourism domain in five Indian languages, viz., Bengali, Hindi, Marathi, Tamil and Telugu, but four more languages, namely, Punjabi, Odiya, Gujarati and Assamese have also been added. An additional UNL based semantic search facility has been provided for Tamil language. A set of ten results is displayed at a time to increase the readability. One has the facility to submit a query by using the Phonetic or Inscript layout, or by using the On-screen layout for Inscript keyboard. Sandhan is able to process the query based on its language and then retrieve results from the respective language. It uses a font transcoder that converts custom fonts into Unicode fonts for processing and generates snippets for each of the retrieved documents that helps the user understand the context of query terms in that document.

The system enables searching Indian language content and, thus, addresses the gap that exists in fulfilling the information needs of the huge Indian population that is familiar conversant with English - which is estimated at 80% of the population. This software is expected to benefit sectors such as academia, tourism and business.

5.3.2 ONLINE SANSKRIT TOOLS

A Consortium of 7 institutes working on 'Development of Sanskrit computational toolkit and Sanskrit-Hindi Machine Translation system' is engaged in developing various computational tools for Sanskrit analysis which assist the reader in understanding Sanskrit texts.

Reading and understanding a Sanskrit Text involves following steps.

- padaccheda
- samasta-pada-paricaya
- Sabda-viSleshaNa
- aakaamkshaa
- anvaya

In the future, it is possible to develop similar readers semi-automatically with the help of following tools. These tools are being developed by the Sanskrit Consortium with financial

support from TDIL programme, MeitY, Govt. of India. Following tools are available on the TDIL portal:

- A simple tool for "transliteration" is available, which provides conversion facility among the Romanised transliteration schemes into Devanagari and also from Devanagari into various Romanised transliteration schemes.
- The Morphological Generator shows the subantas and tingantas (inflectional forms of a noun or a verb). It also handles derivational morphology showing kridanta and taddhita forms.
- Various possible Sabda-viSleshaNas of a Sanskrit word can be obtained from the Morphological Analyszer.
- A tool, "sandhi", is provided to join two Sanskrit words following the Paninian sutras.
- Another tool, sandhi splitter, shows all possible splitting of a given Sanskrit string.

5.3.3 WEB OCR

The development of OCR engines for Indian Languages has been in response to the growth of digital libraries all over the world as they offer access to large number of documents at a faster speed. The OCR process involves converting printed documents into electronic forms using a technology that recognises characters and words from scanned or camera-captured images in a large number of Indian scripts. This content can also be converted electronically for the use of the visually impaired by generating Braille scripts or audiobooks among some possibilities. To aid this purpose, a Web-OCR portal has been made available on the TDIL website, which converts printed documents into electronically accessible format. The area of coverage of the system is Printed Text OCR. A Consortium with IIT Delhi as Consortium Leader is the implementing agency for this portal. The pre-processing modules such as Noise cleaning, Skew detection, binarisation modules have been developed by different consortium institutes. The Language Vertical tasks and integration have been carried out by various consortia members.

The system has been developed and is available for Bangla, Devanagari, Gurumukhi, Kannada Malayalam, Telugu and it will soon be available for Gujrati, Tamil, Oriya, Tibetan, Assamese, Manipuri, Urdu script. It has been developed to support the digitisation of multi-lingual textual images.

It supports file formats BMP, PMG, TIF and a resolution 300 DPI. The file size must not exceed 10MB and the file must in a colour mode of 8 bit grayscale or Black & White.

5.3.4 ANUVADAKSH (ENGLISH TO INDIAN LANGUAGE MACHINE TRANSLATION)

This is a web-based application developed using Machine translation to help overcome the language barrier and encourage language plurality in India. The Ministry of Electronics and Information Technology (MeitY) had come up with the mission of consortia for Machine Translation (MT) Systems and English to Indian Language Machine Translation (EILMT) consortium has been formed as a part of this mission.

Anuvadaksh is a state-of-the-art solution that allows translating the text from English to eight other Indian languages, viz., Hindi, Urdu, Oriya, Bangla, Marathi, Tamil, Gujarati, Bodo. The portal is a collaborative effort of the consortium institutes which have brought forward the integration of four Machine translation technologies: TAG (Tree-Adjoining-Grammar), SMT (Statistical Machine Translation), Analyse and General rule-based (AnalGen), Example based MT (EBMT).

The technical modules, such as Named Entity Recogniser [NER], Word Sense Disambiguation [WSD], Morph synthesiser, Collation & Ranking and Evaluation modules, have been developed by different consortium institutes. The Language Vertical tasks have been carried out by various consortia members.

The engine is available for use for free on the TDIL portal and the input can be upto 1000 words for a single time. The search can be made by selecting a particular domain among general, agriculture, health, tourism for better results in a specific domain.

5.3.5 ANGLA MT SYSTEM (ENGLISH TO INDIAN LANGUAGE MACHINE TRANSLATION)

AnglaMT is a machine-aided translation methodology specifically designed for translating English to Indian languages. AnglaMT is a pattern directed rule based system with context free grammar like structure for English (source language). It generates a `pseudo-target' (Pseudo-Interlingua) applicable to a group of Indian languages (target languages) such as Indo-Aryan family (Hindi, Bangla, Asamiya, Punjabi, Marathi, Oriya, Gujrati, etc.), Dravidian family (Tamil, Telugu, Kannada & Malayalam) and others. It aims to get 90 percent of the translation done by machine and 10 percent is delegated to the human post-editing. It is being developed as a system which can incrementally be trained to handle more complex situations, a uniform mechanism by which translation from English to majority of Indian languages with attachment of appropriate text generator modules as a human-machine interface which can be used to facilitate both its usage and augmentation.

The system can be accessed through the TDIL portal as a web-based application and cannot be used offline. The engine provides a limit of 1000 words in a single input and provides an Inscript keyboard for the source language in which the input is to be given for the text that has to be translated. The application gives more accurate results for specific domains of tourism and health.

5.3.6 SAMPARK (INDIAN LANGUAGE MT SYSTEM)

It is a web-based, multi-part machine translation application on the TDIL portal, which has been developed with the combined efforts of 11 institutions in India under the umbrella project of "Indian to Indian language Machine Translation" (ILMT) funded by the TDIL programme. The ILMT project has developed technology for 9 languages which has enabled machine translation for 18 language pairs. These are: 14 bi-directional pairs between Hindi and Urdu / Punjabi / Telugu / Bengali / Tamil / Marathi / Kannada and 4 bidirectional between Tamil and Malayalam / Telugu.

Sampark uses Computational Paninian Grammar (CPG) approach for analysing language combines it with machine learning. Thus, it uses both traditional rule-based and dictionary-based algorithms with statistical machine learning. Due to the complexity of the NLP system and heterogeneity of the available modules, ILMT system has been developed using Blackboard Architecture to provide interoperability between heterogeneous modules. Hence all the modules operate on a common data representation called Shakti Standard Format (SSF), either in memory or in text stream.

The system is based on an analyse-transfer-generate paradigm during which the source language is analysed and then a transfer of vocabulary and structure to target language is carried out and finally the target language is generated. Each phase consists of multiple "modules" with 13 major ones. Because Indian languages are similar and share grammatical structures, only shallow parsing is done. Transfer grammar component has been kept simple. Domain specific aspects have been handled by building suitable domain dictionaries. The 13 major modules together form a hybrid system that combines rule-based approaches with statistical methods in which the software in essence discovers rules through "training" on text tagged by human language experts.

5.3.7 LPMS (LOCALISATION PROJECT MANAGEMENT SYSTEM)

The LPMS is a web-based portal that provides high end system for daily localisation to small time localisation companies. This system helps facilitate efficient communication with one's surroundings. The LPMS comprises of a system that covers publishing documents for localisation, managing localisation projects, translators and reviewer's dashboard, offline translators' workbench, among others. The LPMS can presently leverage term banks for files such as .txt and .html.[2]

The various components within LPMS are:

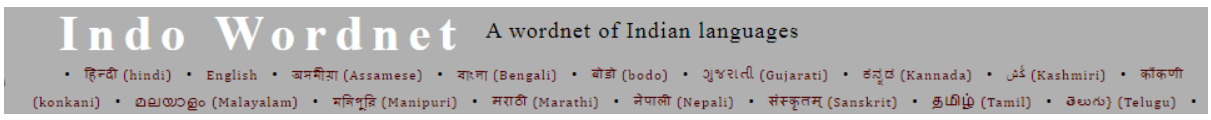
1. Publisher/Client: they are responsible for publishing the localisation jobs to the online community volunteers that include project managers, translators and reviewers, so that they can accomplish their localisation needs.
2. Project Manager: they are responsible for managing the localisation job on behalf of the client/publisher and keep a record of the progress of both the translation and review cycles.
3. Translator: they take part in the job translation cycles and are responsible for supplying translations for community jobs.
4. Reviewer: they participate in the job review cycle and are responsible for approving and reviewing the translations done by translators.

5.3.8 ONLINE HINDI WORDNET

The Online Hindi WordNet is a system that brings together various lexical and semantic connections within Hindi words. It is more than a conventional Hindi dictionary. The Hindi WordNet systematises lexical information according to the similarity of meaning among different words and can be termed as lexicon based on psycholinguistic principles. For each word there is a synonym set, or synset, in the Hindi WorldNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings.

The design for this system was inspired by English WordNet. It uses Unicode for Devanagari fonts and supports viewing on various operating systems like Windows 2000, Windows XP and Windows NT. It can also be accessed through Linux; the viewing, however, may not be as good. This system is also available as an android application.

5.3.9 INDO WORDNET



Indo WordNet is a common lexical database for various major Indian languages such as Assamese, Bangla, Bodo, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Manipuri, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu, which has been created applying the expansion approach from Hindi and English WordNet. This project was a culmination of the

[2] https://tdil-dc.in/index.php?option=com_vertical&parentid=87&lang=en

North East WordNet, Dravidian WordNet and Indradhanush project all of which were funded by the TDIL programme and has been collective endeavour of consortium members.

Every entry in the Indo WordNet comprises of the following aspects along with a related Lexicon in other Indian Languages, including English:

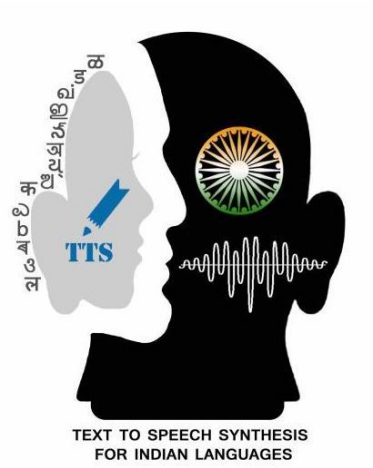
- Synonymy
- Gloss
- Example Sentence

The portal on the website provides three fonts and is available for installation on both Windows and Linux.

5.3.10 SANSKRIT E LEARNING AND MULTIMEDIA

This application service provides a simple and efficient medium for users of all age. A medium from where they can learn the basics of the Sanskrit language. The platform currently offers services to facilitate training in Sanskrit pronunciation, writing, comprehension, grammar, as well as a Sanskrit to Hindi dictionary, and a sandhi generator. The portal also offers an animated story to assist learning among children. This system was funded by the MeitY and was developed by the "Computational Linguistics R&D, Special Centre for Sanskrit Studies, JNU.

5.3.11 TEXT TO SPEECH



This software allows people to interpret machine readable text into Human voice. This is especially conducive for people with visual impairments or reading disabilities as the software allows them to listen to written works on a computer or a mobile device. The Ministry of Electronics and Information Technology, through the TDIL programme, offers various tools, applications and software to enable people to use this technology in different languages. The TDIL portal offers the following applications under its TTS Project:

- -TTS integrated with Screen Reader for Visually Challenged persons: TTS integrated with screen reader are available in Hindi, Bengali, Marathi, Tamil, Telugu, and Malayalam.
- -Browser Plug In: TTS as browser plug ins for 8 Indian Languages (Hindi, Marathi, Tamil, Telugu, Malayalam, Odia, and Guajarati have been developed for Mozilla and Chrome.
- -SMS Reader in Indian Languages (Sandesh Pathak): this SMS reader is available as an android app for 5 Indian Languages – Hindi, Marathi, Tamil, Telugu and Gujarati.
- -Website to keep a check on the voice quality of TTS system: the quality for all languages can be checked and saved from the IIT Madras website at http://www.iitm.ac.in/donlab/hts/

5.4 HOSTING, DISASTER RECOVERY (DR) AND BUSINESS CONTINUITY PLAN

The TDIL website is hosted by the Centre for Development of Advanced Computing or C-DAC at Pune, Maharashtra. C-DAC has been designing, developing and maintaining the various language tools and mobile applications that are ready for public domain on the website for the programme since its launch in November 2010. The centre, since the operationalisation of its data centre at Noida in 2005, has been providing various data services to its clients. It works to provide security to its clients as per international standards. The C-DAC, as a web hosting centre, offers the services like Network and Network security, cloud services, redundancy, and monitoring, among various others, to its clients which are also provided to the TDIL portal. Fig 5.2 is the design of the security system.

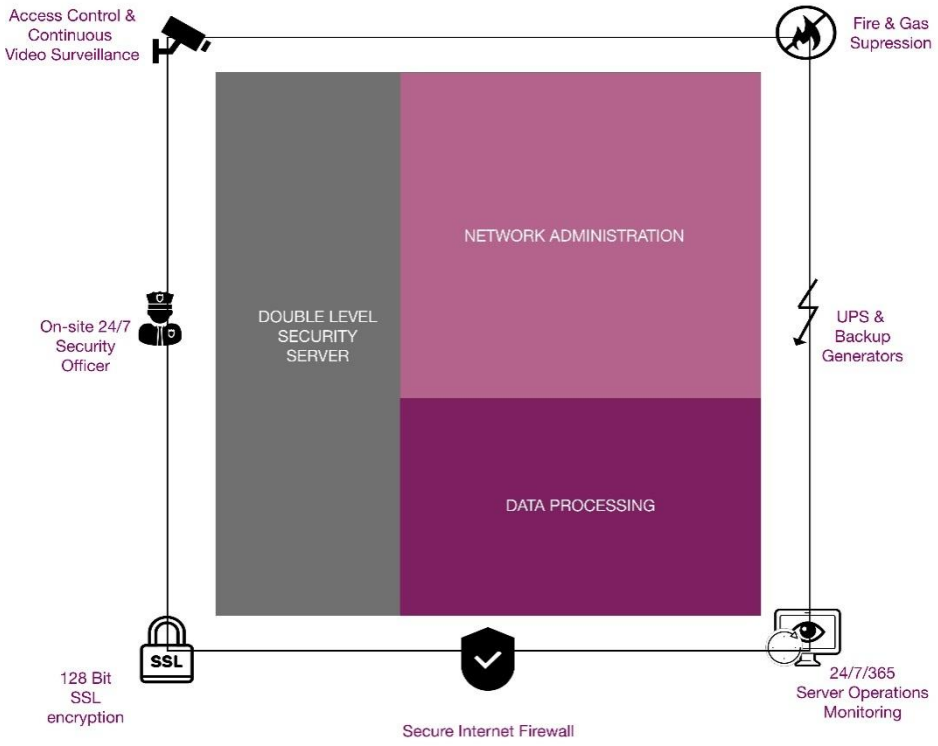


Figure 5. 2: Secure Data Centre

5.4.1 NETWORK & NETWORK SECURITY

The network have been secured by various initiatives. The public access network (internet) links are taken from multiple Service Providers and each network link has been provided through redundant fibres in Ring/Mesh Topologies for guaranteed uptime of 99.8% individually. The BGP enabled network links from multiple service providers assures automatic migration of both incoming and outgoing Network Traffic from one operator to another operator seamlessly, in order to achieve 100% network uptime. The 10gbps backbone network have been configured with redundancy at the cabling level and at equipment level starting from Router, Firewall, UTM, IPS, switches, to avoid any downtime due to equipment failures. The multi-layer network security has been implemented with the perimeter Firewall, IPS and second levels multiple Militarised (Secure) Zones and Demilitarised Zones (DMZ).

5.4.2 CLOUD SERVICES

C-DAC's cloud environment is configured with multiple servers in cluster with common shared storage LUNs (Logical Unit Numbers) from multiple central SAN (Storage Area Network) Storage. The Connectivity from the servers to storage is zoned with multi-pathing and network connectivity from servers to switches is configured with multi-pathing. Hence, the redundancy at the Server level and connectivity level are ensured with the above implementations. Fig 5.2 depicts the interior network design of a data centre.

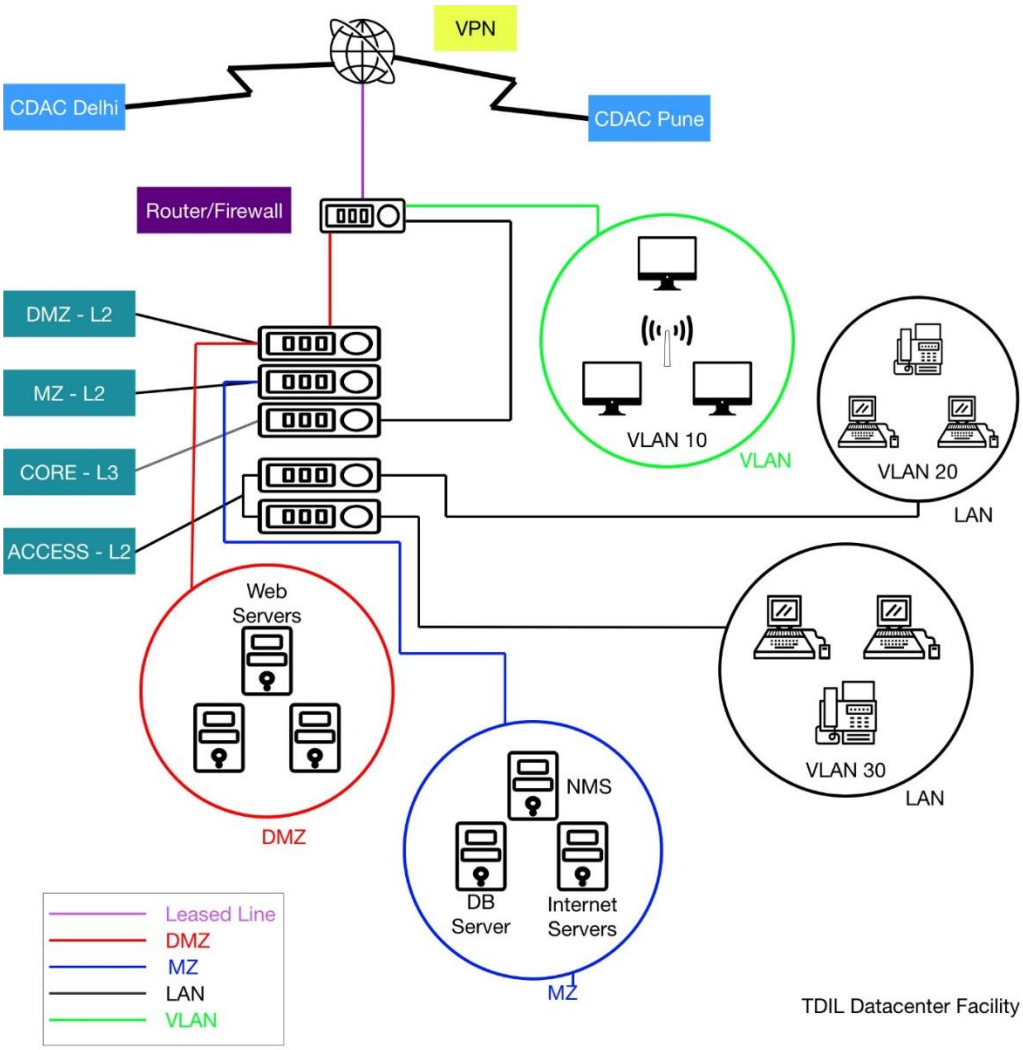


Figure 5. 3: TDIL Data Centre facility

5.4.3 REDUNDANCY

The Applications are deployed in multiple Application Servers and Load Balanced through clustered Load Balancers in High Availability. The Databases are also clustered in High Availability to avoid Single Point of failure. The DNS (Domain Name System) Servers, DHCP (Dynamic Host Configuration Protocol) Servers and e-Mail Servers are all deployed with redundancy. Centralised Repository Servers and Update Servers are used to update the Servers with latest software and patches.

5.4.4 MONITORING

Different types of Monitoring Tools are used to monitor the nodes, servers, resources, services and the performance. The tools generate various alerts based on the configured thresholds at different levels like Normal, Warning, Critical, etc. These alerts are then sent through e-Mail and SMS to the concerned members.

5.4.5 DISASTER RECOVERY

C-DAC Data Centre caters to the data management of TDIL data central portal and it uses Tape library for Data backup to prevent any loss of crucial data in case of any disaster. It is one of the most cost effective ways to store large amounts of data. The DR solution it uses is called C-DAC DRM, which works on a three site architecture of replication solution and online replication and recovery of database data.

C-DAC DRM: It is a centralised DR management and monitoring web application designed to manage DR related activities. It conducts various drill activities like normal copy, reverse normal copy, switchover, switchback, failover, failback, which are fully automated through DRM. One can monitor entire DR setup of various DR deployments through DRM, which reduces manual intervention often prone to human errors and helps in effortless monitoring of entire DR setup. It also has provisions for notifying through SMS/e-mails in case of critical events.

5.4.6 BUSINESS CONTINUITY PLAN

A Business Continuity plan involves the process of creating systems for prevention and recovery to deal with potential threats to the company and to enable ongoing operation before and during the execution of disaster recovery plan. The goal thus is to ensure that in case of a disaster or an undesirable incident, the crucial functions of an organisation do not get interrupted. C-DAC, for this purpose, has developed a traditional data recovery software solution called 'C-DAC Revival' to facilitate business continuity that helps the tools developed and maintained under TDIL to be able to function smoothly despite any contingencies.

C-DAC Revival Solution

It is an end-to-end DR solution which is best suited for data intensive applications like e-Governance and business. It makes use of synchronous and semi-synchronous replication of the database adhering to the goals of zero recovery point objective (RPO) and negligible recovery time objective (RTO). It can be used to manage planned and unplanned outage which enables it to have 24*7 Data availability, thus ensuring long term continuity and availability, performance, productivity and customer satisfaction.

5.5 SECURITY PLAN

The authentication of the servers is done from the central directory servers and for the network devices, it is done from TACACS Servers. The servers' logs are collected in a central SYSLOG server and analysed through a centralised Log Analyser, while for the network devices, it is collected in the TACACS servers. The Firewall, IPS and UTM Logs are also collected in the Centralised Log Analyser Devices / Servers dedicatedly and are analysed on real-time basis. All the servers and network devices are time synchronised with the centralised NTP Server, which

confirms to the coherent and comprehensive log analysis. Security, being the most important pillar of any Data Centre, is established by the process of Risk Assessment, Vulnerability Assessment & Penetration Testing (VAPT) for the Network, Servers and Applications. The Vulnerability Assessment is carried out both manually and with Tools, then Penetration Testing is done for the identified vulnerabilities. The vulnerabilities are reported and patched accordingly. Being a CERT-IN empanelled agency, the complete audit of Web Applications is also carried out in compliance to the OWASP recommendations for the "Safe to Host" Certification and certificate is issued. All the above-mentioned activities, processes, technology implementations and much more are being accomplished by the dedicated, highly skilled and motivated in-house team. The C-DAC Data Centre Team extends its technical support 24*7 and acts proactively to maintain the uptime of 99.982%. The resources are professionally managed to overcome the technology obsolescence and attrition with defined processes[3].

5.6 STATISTICS ON DC

The TDIL portal disseminates many kinds of resources, of which the Publications contribute the most, at almost 60 percent, Localisation makes 1.7 percent, tools like NLP Tools, General tools, Fonts and Mobile Apps contribute around 16 percent, while Linguistic resources occupy around 21 percent of the resources. Among such resources are the tools on this portal, downloadable on free basis and the largest download count has been made by Rupee Symbol Typing Tool with more than 32000 downloads, followed by Unicode Typing Tool with about 18000 downloads and SakalBharti (OTF) with almost 15000 downloads. The count for Text to Speech browser plug-in for Chrome at 13,189 was far greater than the plug-in for Mozilla Browser at 4,563. Other resources can be accessed through online platforms by registering and based on observations extending over a week, the registered users have generally been found to be more engaged than the activated users.

To get an idea, the month and year wise trend of the download count and the user count the period between 2011-2019 were analysed and it was found that the most exponential growth in the download count of resources has been in December of 2017 and July of 2011 while the greatest slump was seen in the month of February of 2018 and August of 2011. While there had been very little rise in the base number of downloads for years between 2010-2014, there is a significant rise since 2015 which again dips in the year 2016 and picks up again in 2017. The download count for January in 2018 reaches a pinnacle of 5000 downloads from around 1500 downloads during the January of 2017. The all year-round consistency has been inflecting very steeply for the last three years- 2017, 2108 and 2019 but the average downloads have very improved significantly. In terms of the user count, the most exponential increase in the registered user count was seen in the month of July in 2011 with an increase to around 1300 from 163 in June that year, followed by July in 2012 with an increased count of 600 from 211 the preceding month. The highest user count in the month of January was registered in the year 2012 and the lowest was registered in September of 2015 at zero registrations. The best performance on the basis of user count was seen in the year 2012 but in general the increase in the number of average registrations has not seen much growth.

[3] https://www.cdac.in/index.aspx?id=pdf_data_centre

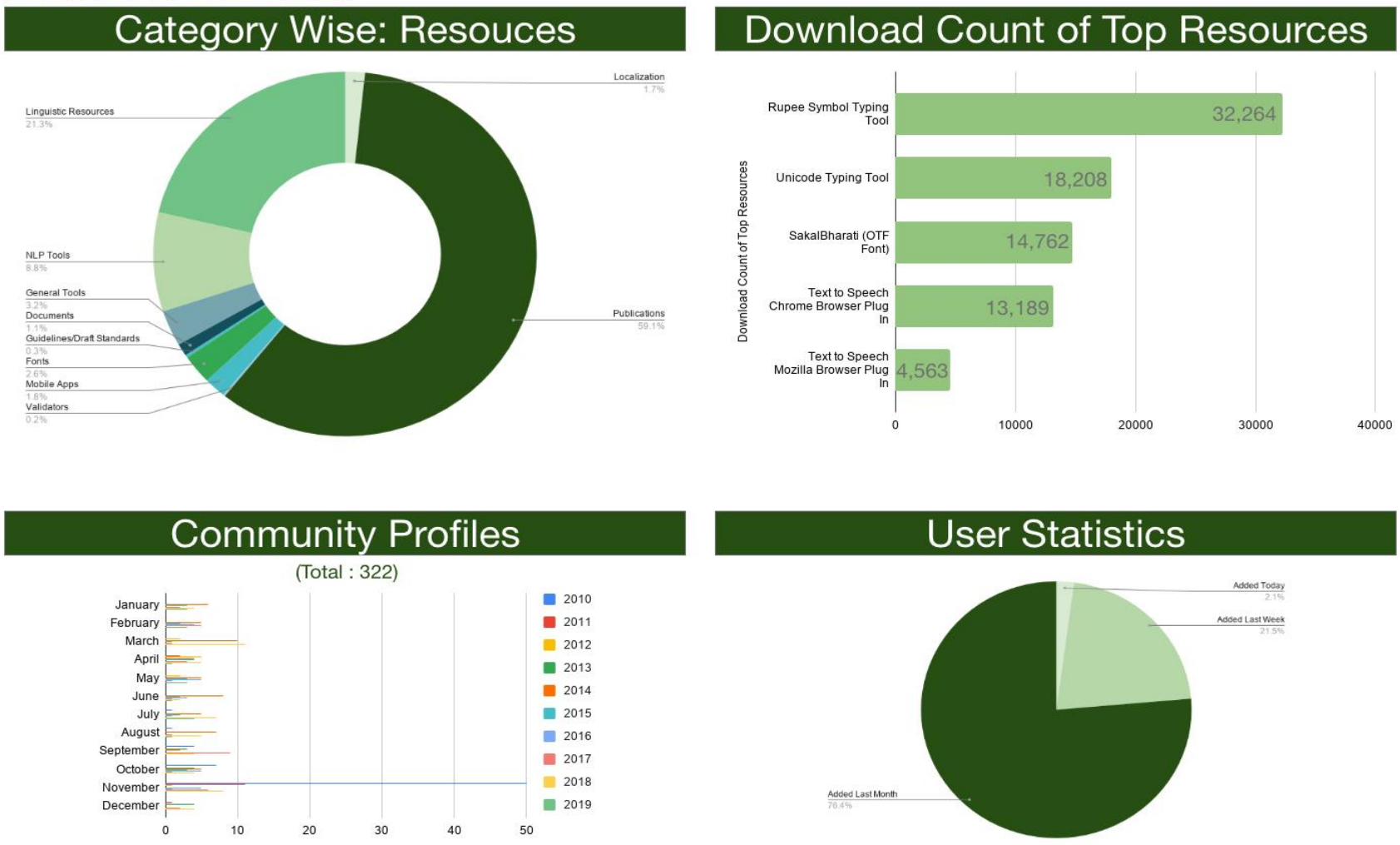


Figure 5. 4: Data Centre Portal resource statistics 1/8

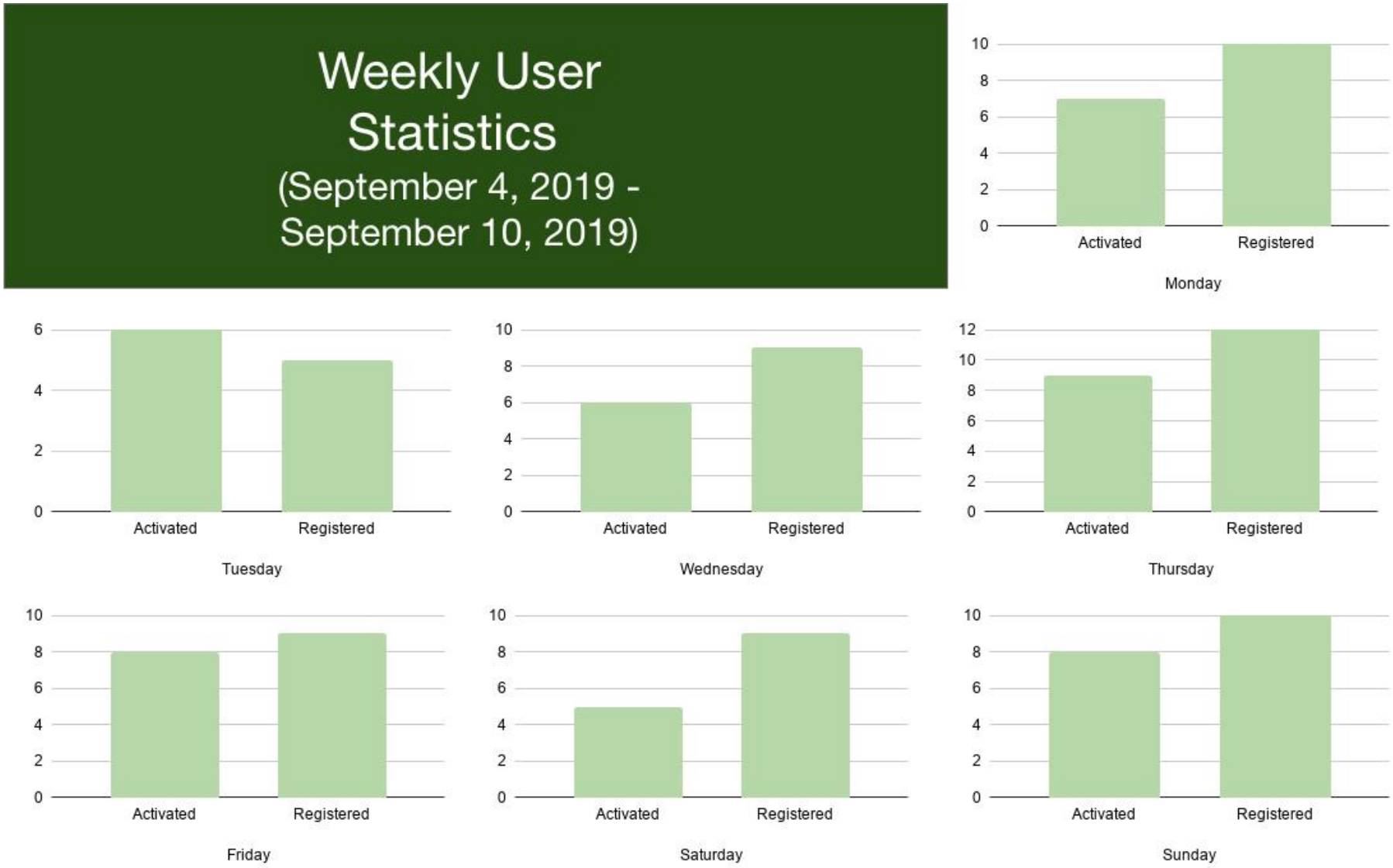


Figure 5. 5: Data Centre Portal resource statistics 2/8

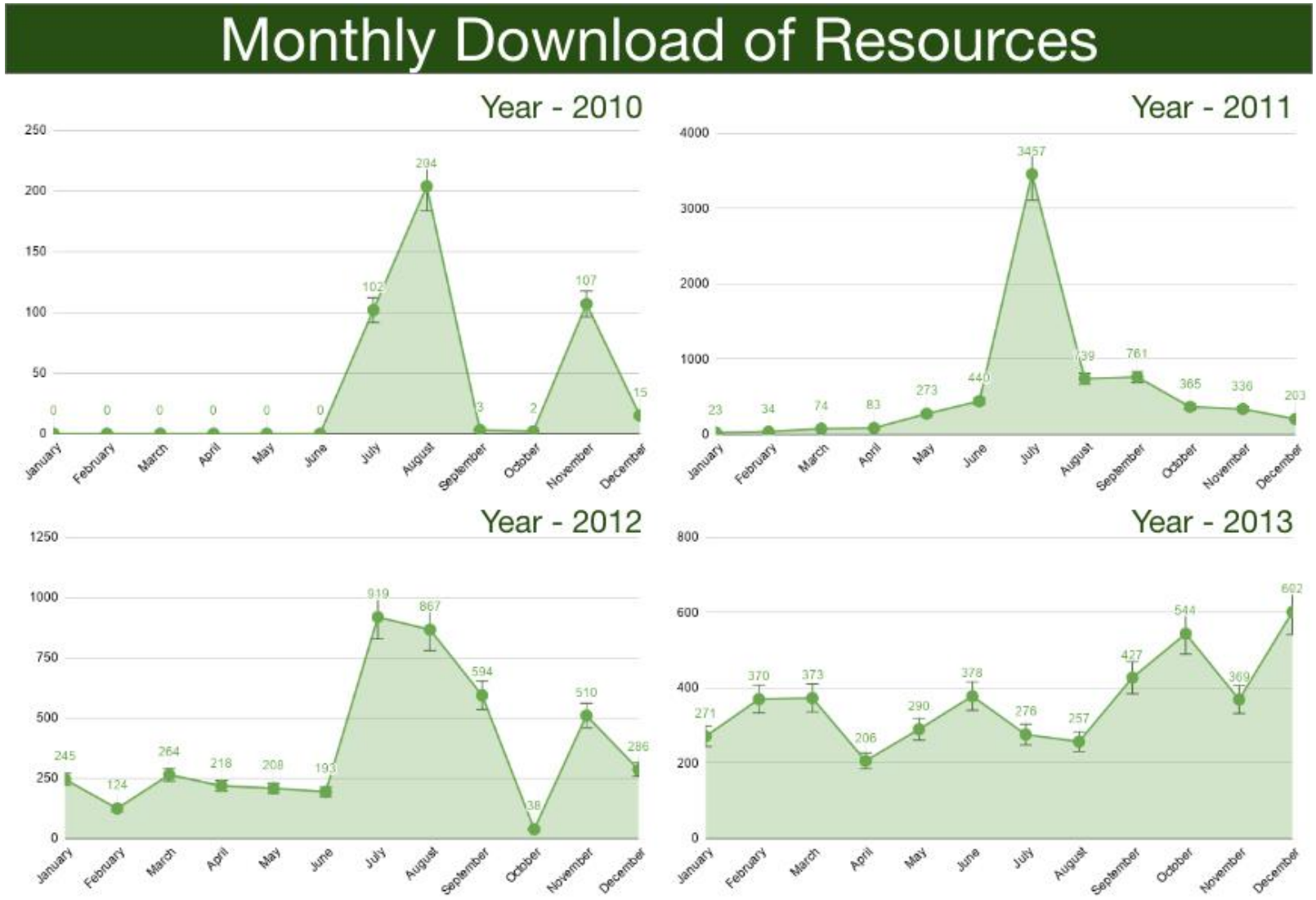


Figure 5. 6: Data Centre Portal resource statistics 3/8

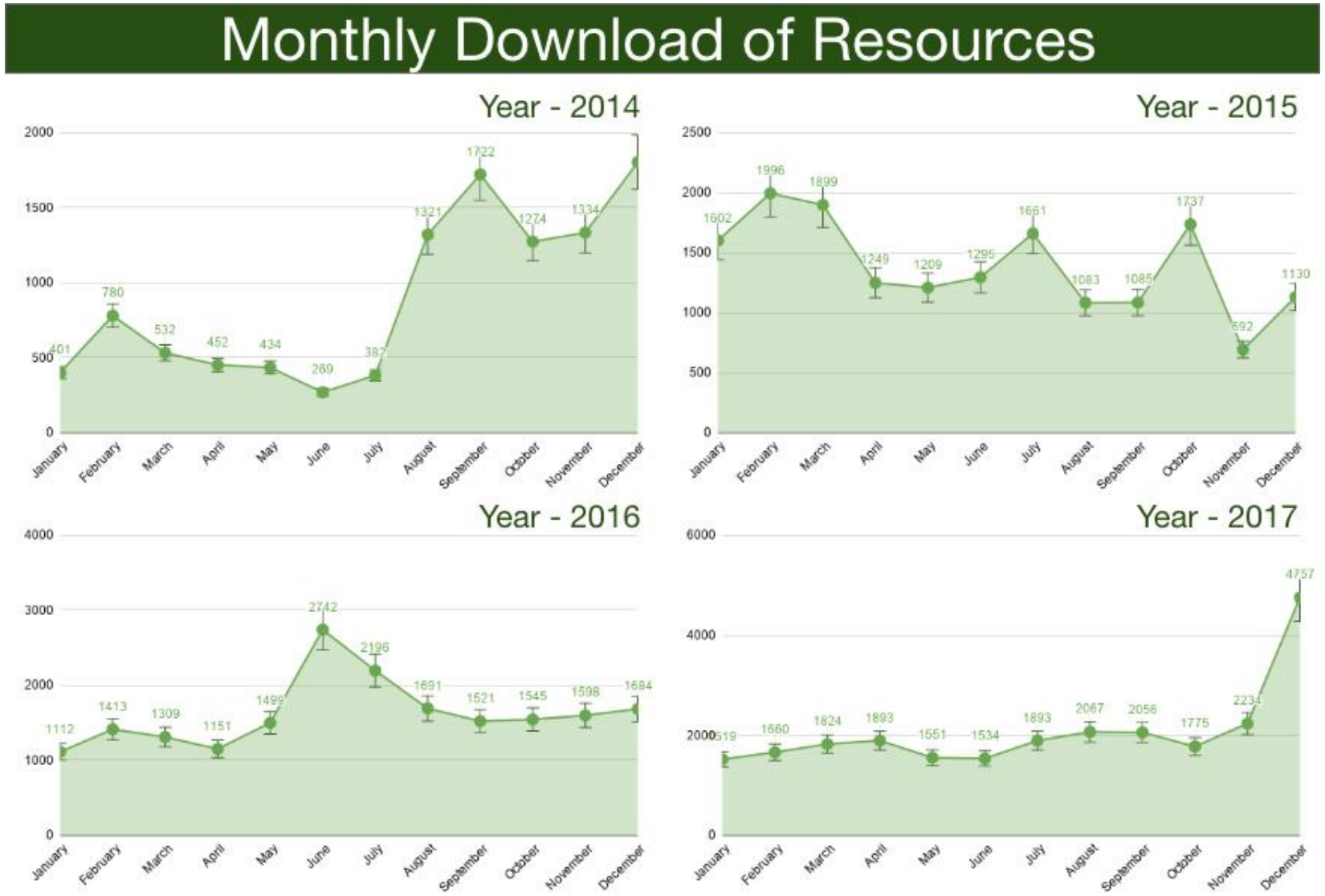


Figure 5. 7: Data Centre Portal resource statistics 4/8

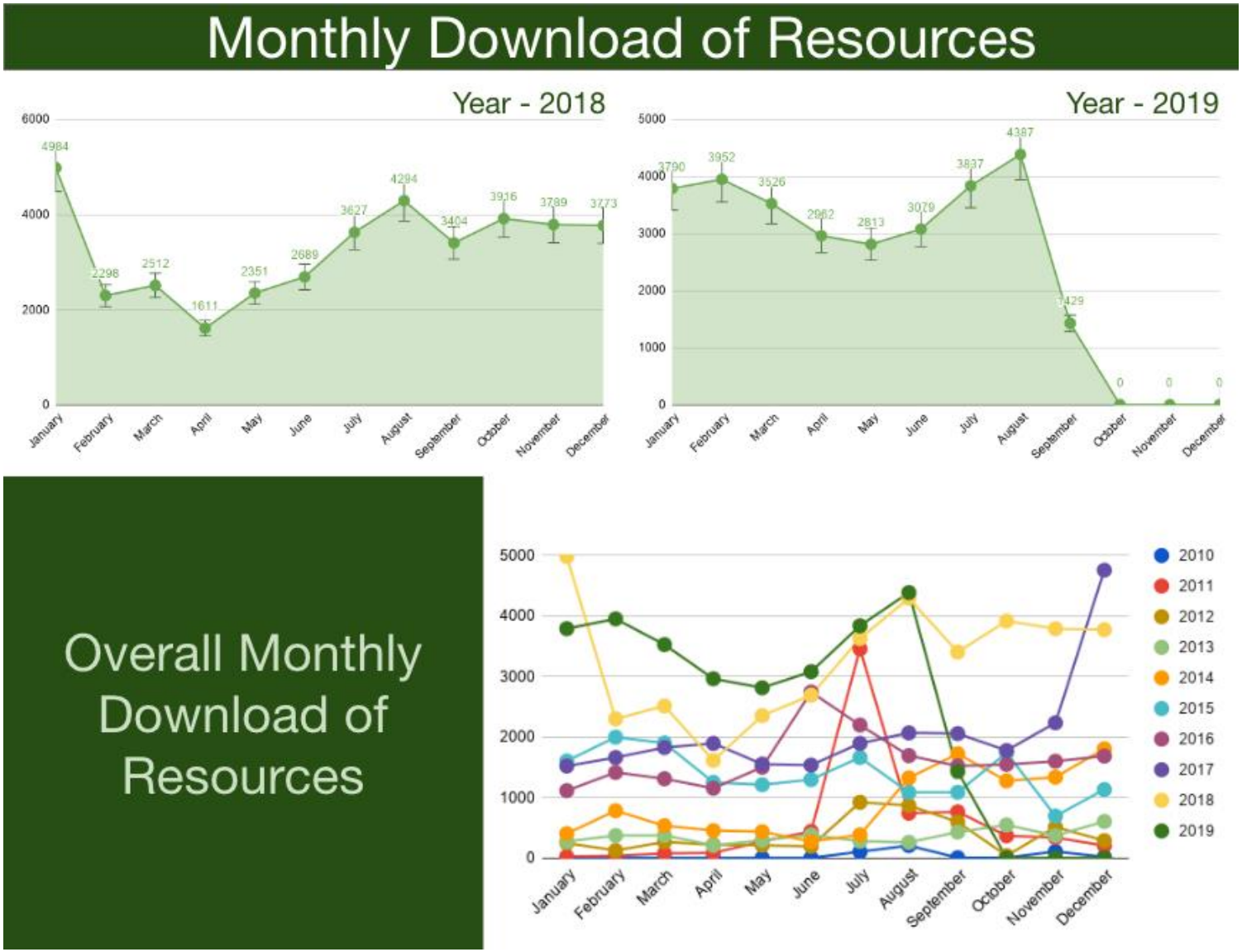


Figure 5. 8: Data Centre Portal resource statistics 5/8

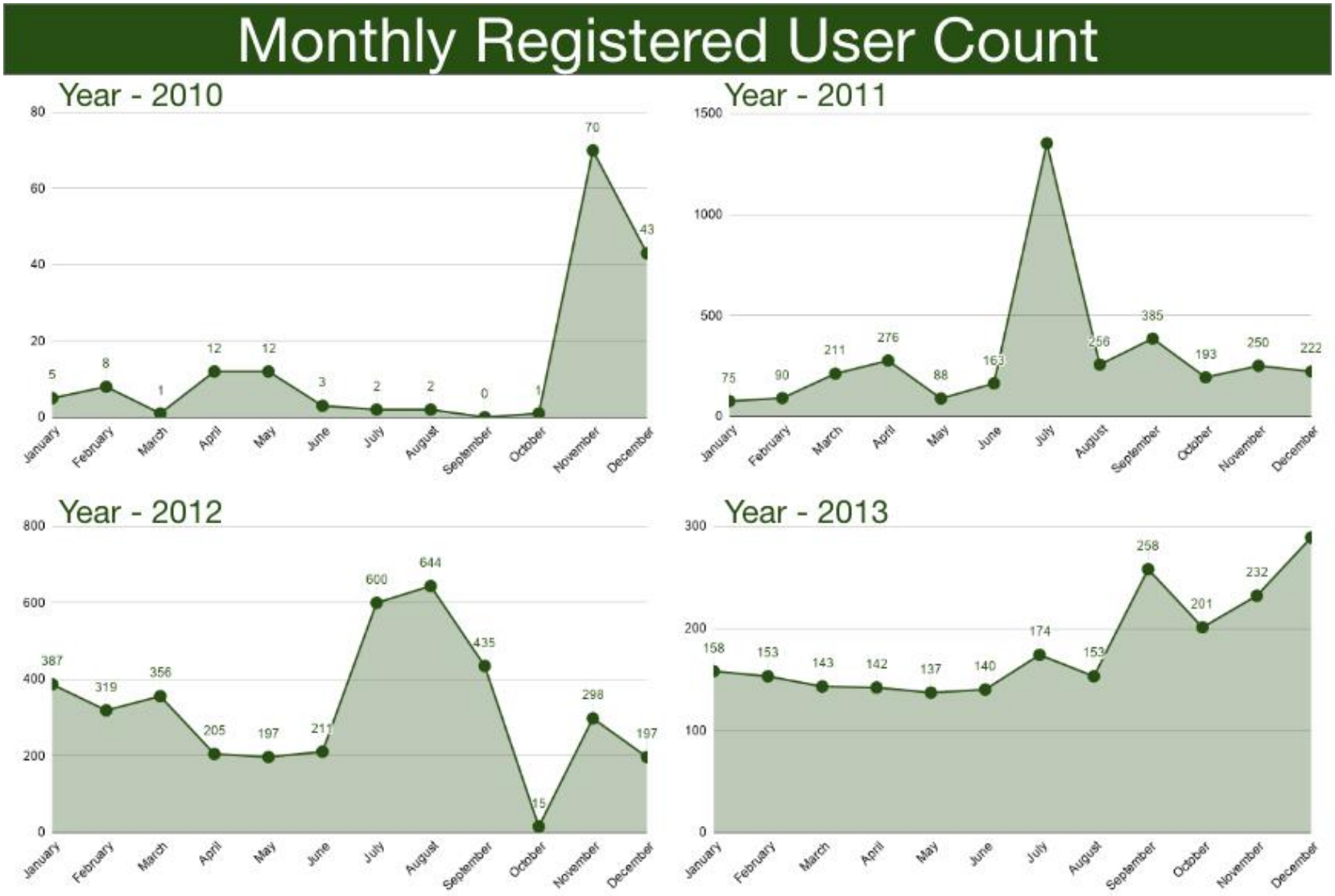


Figure 5. 9: Data Centre Portal resource statistics 6/8

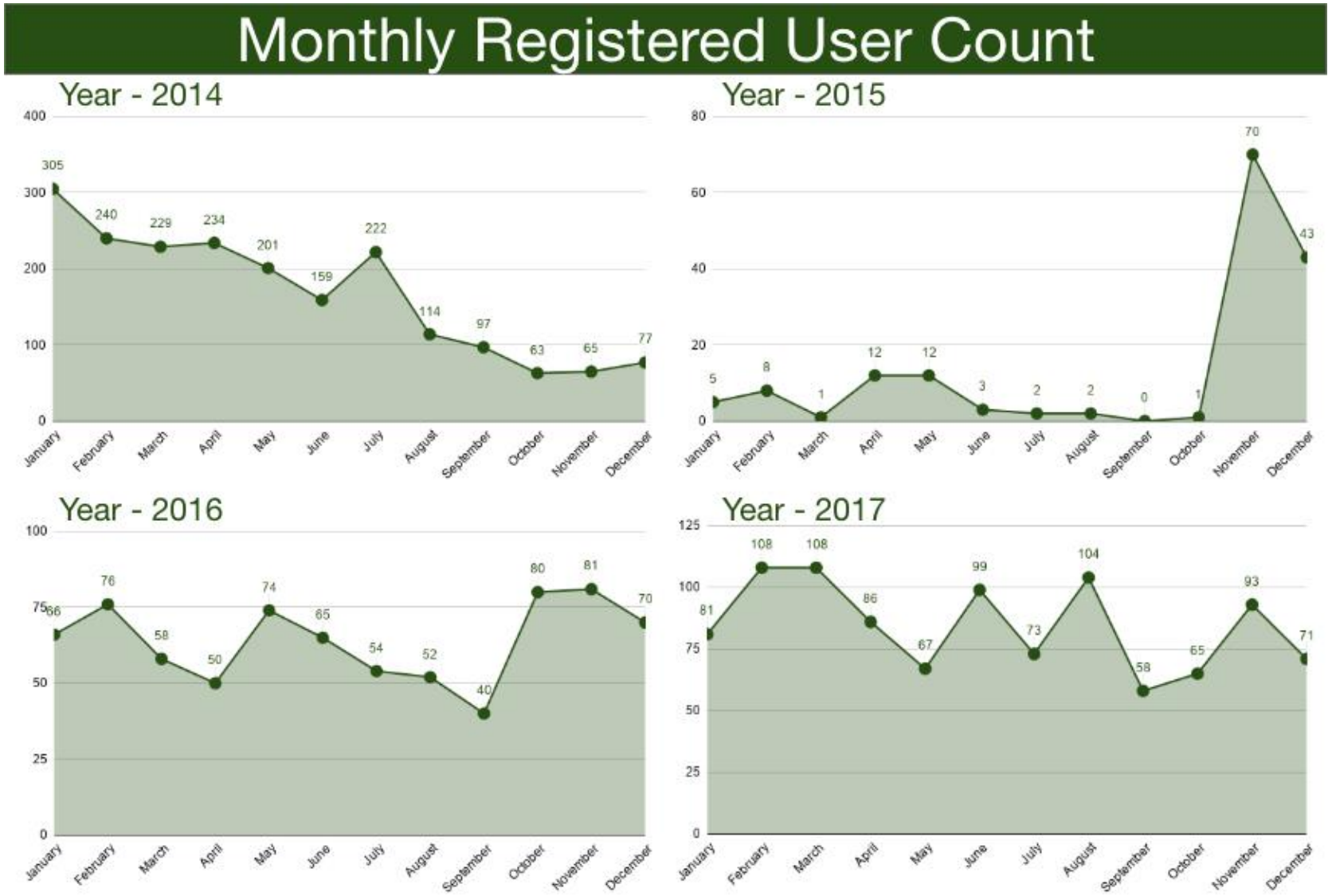


Figure 5. 10: Data Centre Portal resource statistics 7/8

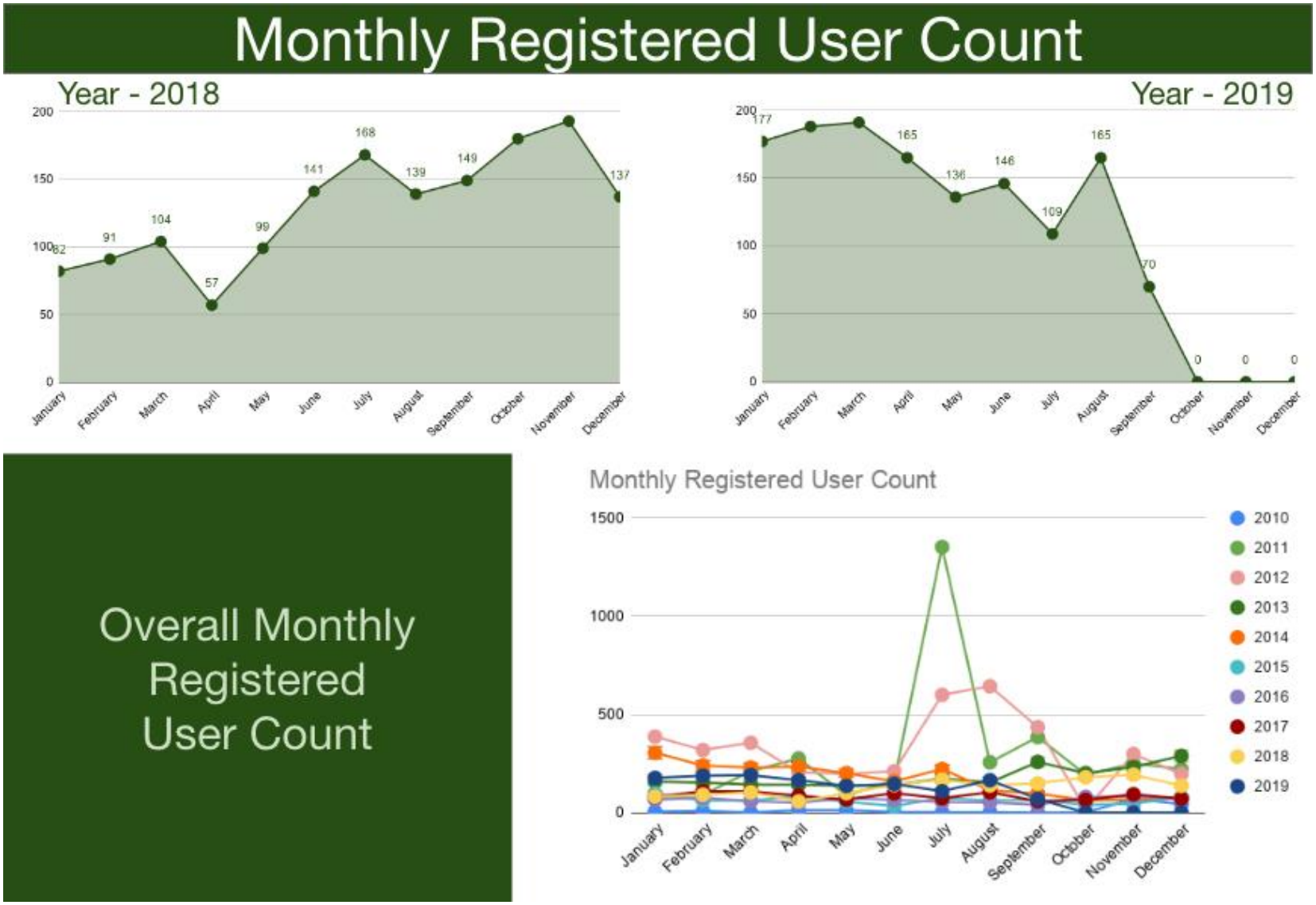


Figure 5. 11: Data Centre Portal resource statistics 8/8

5.7 SCALING UP OF DEPLOYMENT

Find below the Fig 5.12 which is the current decimation of resources and applications through TDIL data centre.

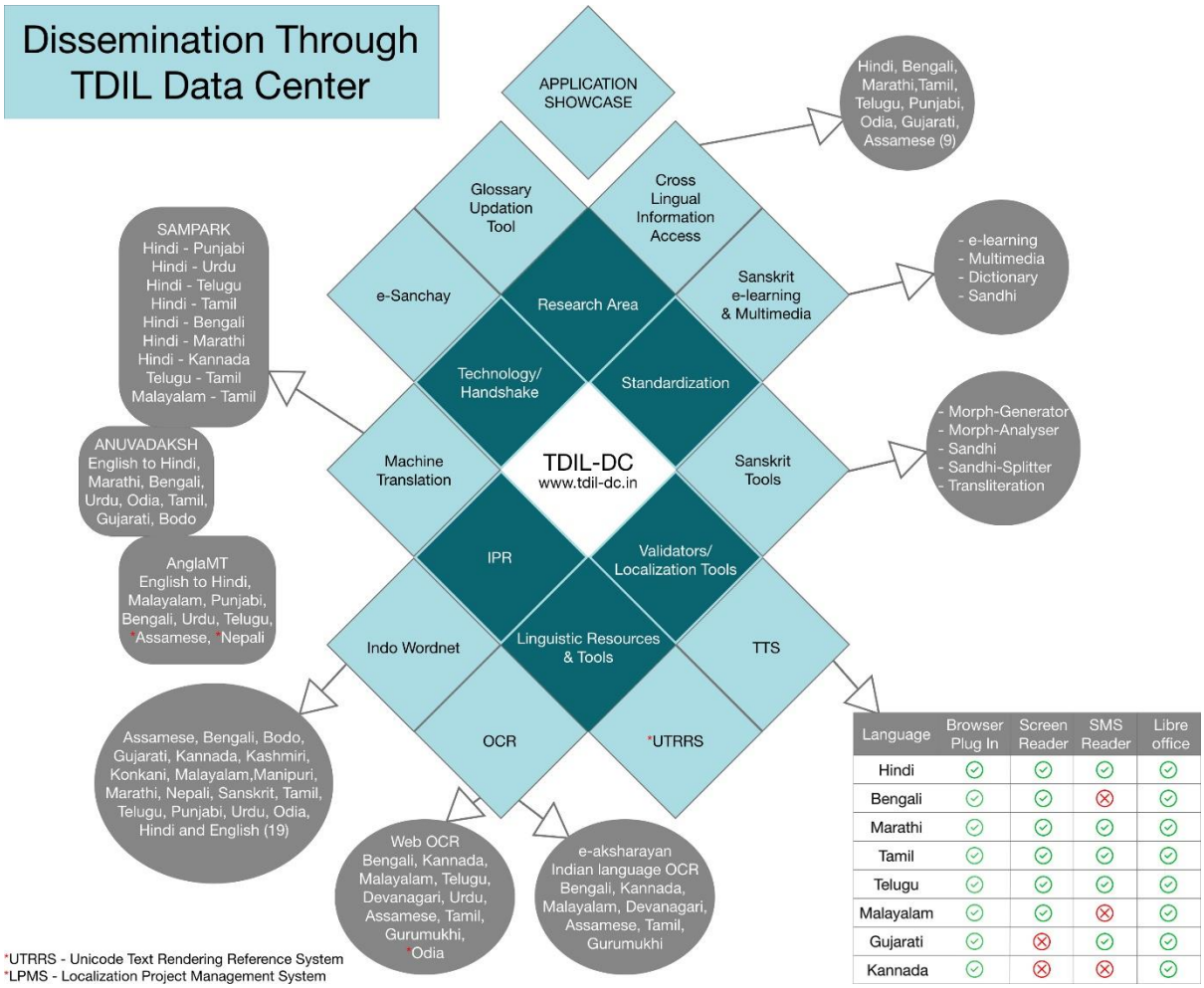


Figure 5. 12: TDIL-DC Portal application architecture

Through this application architecture TDIL is able to deploy and showcase a plethora of applications and technologies developed through its R&D efforts over a period of time. The current deployment, even though commendable, cannot serve the needs of a nation like India. There is need for a systematic scale-up through its own portal, industry, government and start-up ecosystem. A holistic approach for deployment is needed and TDIL should develop a roadmap for deployment in conjunction with emerging technologies.

5.8 RECOGNITION

The TDIL programme has received many prestigious awards and has been recognised for its work through the following felicitations:

- Mahila Prashashan Ratan Sammaan on March 27, 2017
- The Panini Award 2015 for TDIL, MeitY on November 30, 2016

- Manthan Award 2014 under the E-Inclusion & Accessibility Category on December 5, 2014
- The National Council for Promotion of Urdu Language (NCPUL) had felicitated Dr Rajendra Kumar, Mrs Swaran Lata & Mr Mahesh Kulkarni of TDIL on March 10, 2014
- Manthan Award South Asia & Asia Pacific 2013 Winner on December 6, 2013
- MAIT Annual Award on November 25, 2013
- Manthan Award South Asia & Asia Pacific 2012 Finalist- Text To Speech System in Indian Languages (ILTTS) on December 3, 2012
- Manthan Award South Asia & Asia Pacific 2012 Finalist - Development of Tools Technologies & Resources for North East Languages on December 3, 2012
- Skoch Award 2011: National Rollout Plan for all 22 Scheduled Indian Languages on November 22, 2011

Hindi Wordnet- Manthan Award South Asia & Asia Pacific 2009 Winner on December 18, 2009

Case Study: Language CDs

The Indian government has developed a number of language technologies under TDIL. Most of these are available for free use on its online portal, and others may be procured by submitting applications to CDAC Pune. Additionally, some utilities in Indian languages have been compiled and distributed to the public through Language CDs. These CDs, developed under TDIL's National Rollout Plan by MeitY and CDAC, allow users to easily access customised Indian-language programs that can help them in their day-to-day lives. They are a crucial element in TDIL's deployment process, as they aim at reaching the masses who are unfamiliar with either the TDIL website or the process of downloading and installing software.

In order to procure the Language CDs, one must fill in an online application on the TDIL website. The CDs are then delivered, free of cost, by post. They contain installers for the following programs.

1. Bharteeya Open Office: Open Office, a popular alternative to Microsoft Office that offers high quality word, spreadsheet, presentation and drawing tools, has been customised for the Indian audience. It is now available in all 22 national languages of India.

2. UNICODE keyboard drivers for all 22 languages

3. Open Type Fonts in all 22 languages, compliant with UNICODE 6.0

4. The Mozilla Firefox web browser

5. The Mozilla Thunderbird e-mail client, with the Lightning calendar plugin preinstalled

6. GNUCash, an accounting software

7. Inkscape, a graphic design software

8. Tuxpaint, a drawing software designed for children

9. Joomla, a content management software

All of them are open source, and the government encourages users to share them freely. They can be installed on Windows 7, Windows Vista or Windows XP.

5.9 SUMMARY

The TDIL portal is a consolidated database of research papers, articles and so on that outline the structure and details of the programme. It also contains technological resources that can be accessed free of cost by the public:

- Sandhan
- Online Sanskrit tools
- e-Aksharayan
- Anuvadaksh
- AnglaMT
- Sampark
- Localisation Project Management System (LPMS)
- Online Hindi Wordnet
- Indo Wordnet
- Text-to-speech software
- Sanskrit e-learning and multimedia

The website is hosted by CDAC Pune, the institution in charge of approving and deploying TDIL technologies. CDAC Pune oversees network security, cloud services, redundancy, server monitoring, disaster recovery, and the formulation of business continuity plans. Statistics show that the site is accessed by a large number of users every day.

Chapter 6: Understanding TDIL Commercialisation Efforts

6.1 OVERVIEW

Several innovations in language technology, from machine translation to automatic speech recognition, have come into use, especially over the last two decades. Some of these were driven directly by government investment. Others made their entry via the global ICT market. The diffusion and transfer of new language technologies within India is certainly worth study, as is their commercialisation by different private and public players. This section starts by outlining the history of language technologies in the Indian context, providing brief descriptions of the most prominent ones and discussing the transfer of these technologies between parties. Next, it focuses on how they were incorporated in commercial products and services. Finally, it examines how language technologies have become entrenched in different sectors.

6.2 INTRODUCTION

India is the top exporter of ICT services in the world, with the industry generating a market value of $126 billion in 2018 alone[4]. Cities like Mumbai and Bengaluru have emerged as hubs of information technology, attracting multinational conglomerates and local start-ups alike. One might be forgiven for marvelling at the rapid expansion of the ICT sector in the country. Upon examining the statistics, however, one finds that there has in fact been inadequate focus on the development of the *domestic* ICT industry, which in 2018 generated a mere $41 billion. According to a UN report, India ranks 109th on the use of ICT, and the online service of the Indian government was rated 33rd internationally (Fig. 6.1) [5]. A major reason for these statistics is poor accessibility of e-resources to speakers of most Indic languages.

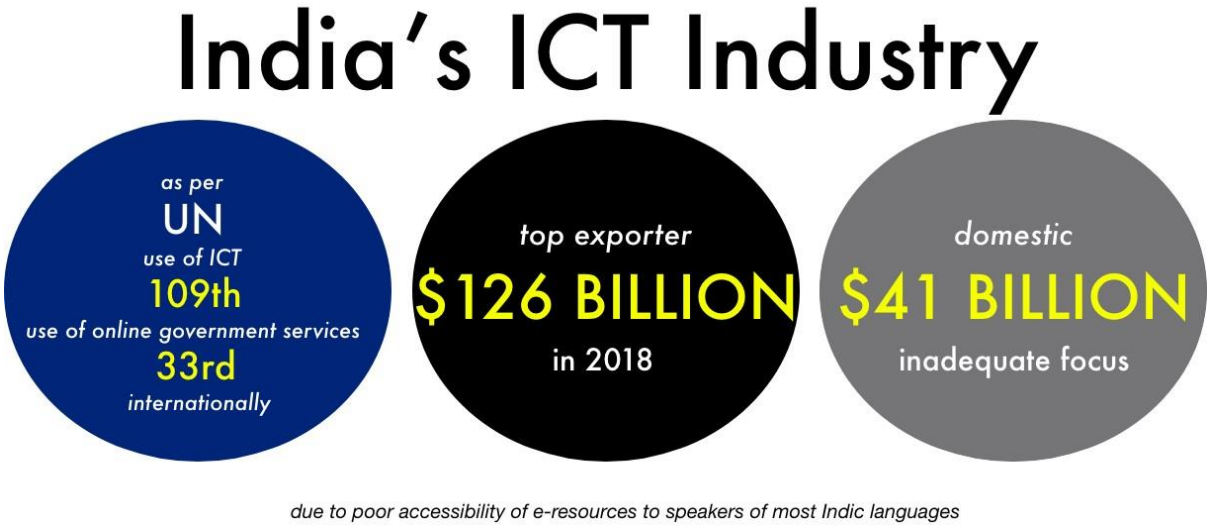


Figure 6. 1: ICT industry in India

Although the ICT revolution has brought smartphones and internet providers to every corner of India, the bulk of e-services and products have remained out of reach for the country's masses. This is because they are designed for fluent speakers of either English or Hindi, both of whom constitute numerical minorities in India. In order to truly digitise the country, it is necessary to make ICT available in each of India's 22 major languages. Only then can the government communicate directly with the average citizen and gather reliable data at the grassroot level. The role of language technology in accomplishing this goal cannot be reiterated enough.

[4] IT industry in India – Statistics & Facts (2019). In *Statista.* Retrieved from https://www.statista.com/topics/2256/it-industry-in-india/

[5] India ranked top exporter of ICT services: UN report (2017, June 16). *The Hindu.* Retrieved from https://www.thehindu.com/news/national/india-ranked-top-exporter-of-ict-services-un-report/article19086441.ece

6.3 HISTORY OF LANGUAGE TECHNOLOGY IN INDIA

Up until the early phase of the TDIL framework (1985-91), computerisation in India was low. Government information was largely conveyed through television and radio broadcasts. There was little to no market demand for language technology. Post liberalisation (1991), as India opened up her markets to international trade and investment, the importance of evolving better communication technologies became apparent. In particular, it was evident that unless there was investment in language technology, the digital revolution would be unavailable for the majority of Indians. Moreover, it would severely inhibit the political awareness of citizens and hence hamper effective governance.

It was against this backdrop that, in 1991, that the Department of Electronics launched the Technology Development for Indian Languages (TDIL) programme. As part of this, several technologies have since been imported and developed to facilitate translation between English and the 22 major Indic languages. This process has seen the involvement of the government, start-ups and major e-business houses across the world.

6.3.1 NLP AND ISCII

Natural Language Processing (NLP) is the process by which text in any given language is interpreted by a computer. The first step in this process is transliteration, or the conversion of the script into an encoding standard that is easily understood by the computer. Between 1986-88, the Department of Electronics developed a standard to represent Indic scripts on computers. This was known as the Indian Standard Code for Information Interchange (ISCII). It was formally adopted by the Bureau of Indian Standards in 1991, and was since outsourced to IBM, Apple and other multinational companies. To this day, ISCII remains a significant influencer of the information revolution in India. When India became a member of the Unicode Consortium, ISCII was abandoned; however, the encoding of Indic consonants and vowels in the Unicode system is based largely on their old ISCII codes [6].

Any platform that recognises Unicode will render Indic scripts, whether it is a web browser (such as Mozilla Firefox) or a text editor (e.g. Microsoft Office).

6.3.2 MACHINE TRANSLATION

Not only are English and Hindi treated as bridge languages in most parts of the country, but they are also privileged over all others in that they are classified as India's official languages. All the work of the central government is conducted in either English or Hindi. This means that for the majority of Indian citizens, the languages of governance are not easily accessible. Moreover, the fruits of the digital revolution have, for the most part, been reaped by the English-speaking minority, as e-resources in Indic languages are relatively scarce. Translation, between English and different Indian languages, has therefore been recognised as an imperative goal by successive governments. Ever since machine translation became accepted as feasible in the West, India has had her eyes on the technology.

---

[6] ISCII (Indian Standard Code for Information Interchange). In *CDAC* website. Retrieved from https://cdac.in/index.aspx?id=mlc_gist_iscii

As has been detailed elsewhere in this report, machine translation in India began with the MAT, MATRA and MANTRA projects. Early Indian efforts (c. 1986) in the area of machine translation were pioneered by public research institutions. These institutions included:

- Indian Institute of Technology (IIT) Kanpur
- Computer and Information Sciences Dept, University of Hyderabad
- National Centre for Software Technology - now known as the Centre for Development of Advanced Computing (CDAC), Mumbai
- Centre for Development of Advanced Computing, Pune

Since then, private entities have also become involved in machine translation for Indian languages. One such entity is Microsoft Research, which is working on the rule-statistical and hybrid translation methods. Supersoft India Private Limited and the IBM India Research Laboratory are other drivers of research in the field.

The machine translation technologies developed by central research teams have been used by start-ups, small enterprises and state governments. Some major clients include the Govt. of Kerala, Govt. of Andhra Pradesh, and the National Institute of Open Schooling (Noida). Most recently, the government has been transferring its MT technology to a start-up called e-Bhasha Setu Services Pvt Ltd.

The developed MT systems are also available on the TDIL portal.

6.3.3 AUTOMATIC SPEECH RECOGNITION

Since it is not possible to write (or type) faster than the speed at which someone is talking, verbatim transcription is not a viable option in most fields. Moreover, since human energy and productivity decline over time, the risk of error increases with the length of the segment being manually transcribed. Speech recognition technology, on the other hand, produces nearly instantaneous output with no margin for human error. While it is far from perfect at the moment, further research in the area should yield promising results - especially if it is made in tandem with larger research institutes across the world.

Automatic speech recognition in India was developed only in the late 1990s by the Indian government, although private companies had been working on it for two decades prior to it.. The first major public project in this area was undertaken in 2010, when a team led by Prof S Umesh from IIT Madras worked on creating speech-based access for agricultural commodities in six Indian languages. This system made it possible for farmers to investigate the location-based market prices of their produce in real time. Additionally, it could be used to ask for weather forecasts. All the user had to do was call a helpline and provide some details to a computer. This computer would then transcribe their voice commands, interpret them by matching the words with a dictionary, and then return the appropriate audio outputs.

This project was expanded in 2014, and currently provides the same services in 9 Indian languages. Its services have been deployed in 12 states.

6.3.4 CROSS LINGUAL INFORMATION ACCESS

Just over half of the most popular websites on the World Wide Web are in English, and even the second most used language on the internet - Russian - can be found in merely 6% of the top

websites. Obviously, accessing information from a website that is in a language other than one's native language is very difficult, especially for speakers of non-European languages. Cross-language information retrieval, or the process by which queries in one language are matched with web results in other languages, is a potential solution to this problem.

In 2010, the Government of India launched the Cross-Lingual Information Access (CLIA) project, centred around a team of linguists, computer experts and academic researchers. It aimed at making it possible for the common Indian citizen to easily access the bulk of the internet. The end result of the project, known as Sandhan, is a free mobile application and web crawler that relies on machine translation to access tourism-related content. It hopes to allow Indian travellers to easily access information on transport, tourist attractions, amenities and so on from government websites and Wikipedia.

6.3.5 OCR AND OHWR

Despite the fact that we live in the era of smartphones and other digital technology, much of our written material continues to exist in hard form. Printed documents, sign boards, images and so on are still the norm. In several situations it becomes necessary to extract data from printed content - whether it is to analyse said data or reproduce it in other media. The technology that enables this is optical character recognition (OCR).

In 2007, as part of the TDIL programme, IIT Delhi and thirteen partner institutes created the e-Aksharayan, the first OCR system to work with Indian languages. As of 2019, this software, available for download on the TDIL website, is able to scan printed documents in 7 Indian languages and return editable text based on the original input. It has functionality for multiple fonts and layouts and can recognise large documents. In addition to the e-Aksharayan, the government has also released a simpler web OCR system that can be accessed from any browser. This software supports documents in six Indian languages, and works with a much more limited file size.

Online handwriting recognition (OHWR) is another technology that scans and interprets documents. However, it is specifically designed to work with handwriting. It was first launched in 2006 by researchers from the Indian Institute of Sciences (Bangalore) and the Centre for Development of Advanced Computing. As of 2019, the full software recognises handwriting in six languages, and the demo version (available on the TDIL website) offers free services in two - Hindi and Tamil.

6.3.6 TEXT TO SPEECH

The conversion of written text to audio files that can be played by a computer is known as text-to-speech technology. First developed in Japan in the late 60s, TTS has become a common feature in most gadgets today, from smartphones to laptops. As can be imagined, it is extremely helpful for the visually challenged and the illiterate. However, TTS can also be used to generate automated responses, translate conversations between speakers of different languages, teach new languages and perform a host of other functions that we are only beginning to conceptualise.

The first public TTS project was launched in 2009 by a team led by Prof Hema Murthy of IIT Madras. A larger initiative was undertaken when, three years later, the Department of Text-to-

Speech Systems in Indian Languages was created. So far, five major TTS systems have been made available under TDIL:

- A text-to-speech system with a screen reader that recognises text in six Indian languages.
- A browser plugin for Mozilla Firefox and Google Chrome that recognises eight Indian languages. This has been utilised by the Mahir Katha (Govt of West Bengal) for providing agricultural information, National Bank for Agriculture and Rural Development (Mumbai), and the Vikaspedia portal.
- A TTS plugin for Libre Office.
- An Android app (*SMS Reader*) that recognises 5 Indian languages.
- A web page that provides basic TTS features in 13 languages.

Additionally, TTS functionality in 9 Indian languages has been incorporated in Indus OS, an Android-based operating system that is used by eight mobile manufacturers in India. According to a TDIL briefing, the Indian government expects its partnership with OS Labs India Pvt Ltd to connect an additional 300 million consumers to the internet in the next five years - and the inclusion of Indian languages in Indus OS was likely a key influence in this estimation.

6.3.7 TEXT CORPUS

In order to perform the kind of statistical analysis that is required for machine translation, text-to-speech and other language technologies, it is necessary to have a bulk of raw linguistic data in written form. This data, known as the text corpus, may comprise different kinds of texts, from children's storybooks to transcript conversations.

The Indian government has been amassing such data since Feb, 2009, when the Indian Languages Corpora Initiative was launched under the leadership of Dr Girish Nath Jha, JNU. This aims to collect and annotate texts in 12 Indian languages, including English, from the areas of tourism and health. The source language used is Hindi. It is expected that the project will generate ample linguistic data to improve the accuracy of assorted language technologies.

In phase I the focus was on Health and Tourism domain. In phase II two more domain were identified Agriculture and Entertainment. Apart from these domains 16 other domains were identified for monolingual corpus collection. These domains were Art & Culture, Economy, Health, History, Philosophy, Politics and Public Administration, Religion, Science & Technology, Education, Sports, Tourism, National Security and Defence, Law, Society and Community, Literature, and Geography. Total 18 domains were collected with 260 sub-domains.

6.4 RECOMMENDATIONS

The previous section has provided an overview of the most prominent language technologies currently in use. It has also pinpointed cases of the developed technology being freely transferred to other parties. This section looks at the dissemination and commercialisation of the above-mentioned technologies and makes recommendations on how to improve these processes.

6.4.1 AVAILABILITY OF INFORMATION

While preparing this report we found that there were many online research papers, articles and websites that contained information on TDIL. First and foremost, we recommend that all this data be compiled and uploaded onto the official TDIL website (http://tdil-dc.in).

The site's home page offers several links that might be useful to someone familiar with language technology but may not reach the layperson. In order to work around this, the home page could contain (a) a brief history of language technology, (b) an introduction to TDIL, and (c) a list of real-life applications of language technology in India. This would make the site more accessible to people from non-technical backgrounds.

Each section of the website contains valuable information. However, adding more statistics, background information and links to external sites would make it even more accessible to users. It would also increase the chances of the TDIL website appearing in search engine results.

To illustrate this: in the page 'Technology Handshakes' (http://tdil-dc.in/index.php?option=com_vertical&parentid=29&lang=en&Itemid=520), for instance, there could be a section that answers questions such as:

- Which entities has the government outsourced its technology to?
- What has been the impact of each technology on e-commerce?
- Which private researchers have worked with the government for technology development and/or testing?
- How has the government spread awareness on the developed technologies?

Answers to these questions may be found after searching through different sections of the website. The point that must be noted, though, is that many may not have the familiarity with the internet to do so. In order to hold the attention of the public, the site admin could hire more content writers, consolidate more TDIL data, and share it with the public in simple language. We also recommend that the graphic designers make the visual interface more minimalistic, so as to encourage new visitors to explore the site at their own pace.

We are pleased to note that an online search for phrases such as "tdil" yields multiple results from popular news journals and articles. However, the TDIL needs to take more steps in order to increase awareness among people who do not speak English or have access to the internet. The developments in India's ICT industry need to be communicated to the masses in a way that they can understand. The research team was informed that the TDIL is already working on this.

6.4.2 DEVELOPMENT OF START-UP BUSINESS ENVIRONMENT FOR INDIAN LANGUAGE TECHNOLOGIES

Each of the technologies developed under TDIL is ready for usage, although some are in a higher stage of functionality than others. The process by which they may be procured is straightforward: first, the applicant, whether an individual or an entity, submits a request to one of the affiliated research institutes, or through the TDIL website. Then the request is processed

through the Centre for Advanced Computing, Pune and the technologies are provided to the requester.

According to an official memorandum submitted on the 24th of May 2019, the linguistic resources developed under TDIL are now available free of cost to all Indian academic researchers and start-ups, although they are required to sign non-disclosure agreements with CDAC Pune. Other entities that wish to use the developed technology must bear a fraction of the cost of developing it. This way, the government provides a boost to small ICT enterprises and simultaneously popularises TDIL among the academic community.

The rationale behind making the TDIL resources so easily available to private entities (who will likely incorporate them in larger commercial technology packages) is that more e-products will be created as a result. The more goods and services that spring up, the higher the GDP share of the ICT sector becomes. It is widely expected that this in turn will create more job opportunities and make the internet more accessible to the non-Anglicised citizens of India.

However, expanding the e-market may not necessarily lead to the democratisation of information. If the primary goal of the government is to make the internet available to the masses, it should not only carefully monitor the commercialisation of language technology by the private sector, but also find independent ways to utilise the valuable tools and resources its own researchers have created.

For instance, consider its machine translation systems. Machine translation, when used together with speech recognition and text-to-speech technologies, has great potential for facilitating communication between the speakers of different languages. This may be achieved as follows. Assume that one of the communicants involved is a native speaker of Hindi, and that the other is a native speaker of Manipuri. When the Hindi speaker says something, his speech is immediately run through ASR technology, which essentially transforms audio into text. This text is then machine translated to Manipuri. Finally, the translated text is read out loud by the integrated TTS programme. The entire process could be achieved through a single commercial software.

As of now, such a software is more likely to be developed by a private company, which, unlike the government, is usually more interested in generating profit than in providing affordable services to the public. This is because the primary obligation of any private venture must be to its shareholders. We are therefore of the view that the developed software is likely to be available to only a minority - at least in the early stages. On the other hand, if the government were to take a more active role in integrating its technical resources, it could provide the same software at negligible costs, if not for free. Not only would this make the technology more available to the public, but it would also increase public faith in government efficiency and expertise.

Of course, as things stand, large corporations have far more years of technical expertise than the government. In order to balance this out, we recommend that the government enter into MOUs for research with technical experts in both foreign governments as well as private players. We also recommend that it support the smaller start-ups that work with language technology.

6.5 EXISTING VERNACULAR PRODUCTS AND SERVICES FOR DIFFERENT SECTOR

Although we are yet to actualise the potential of language technology in India, we have already seen several initiatives in that direction within the ICT market. The most creative and inspiring stories are those of start-ups. These shed light on how scientific knowledge, if made easily available to small entrepreneurs, can increase innovation and simultaneously reach a large number of clients. This section briefly looks at start-ups that have used language technology in different sectors.

6.5.1 TECHNOLOGY SECTOR

IndusOS

Initially known as FirsTouch, Indus OS is an operating system founded in 2015 that brings the smartphone experience to speakers of Indian languages. Not only does it provide keyboards and text-to-speech functionality in 12 Indian languages, but it also contains an option to translate between English and any Indian language with a single swipe. Its appstore, AppBazaar, allows users to download a number of applications in their native languages. Within 18 months from the day it was launched, Indus OS had become the second most used operating system in India, reaching over 10 million users. It has partnered with 10 phone manufacturers and is available on 100s of devices.

Vernacular

Vernacular is the start-up that created Vernacular Intelligent Voice Assistant (VIVA), a voice assistant software that provides services for businesses in the banking, insurance and food service sectors. It recognises speech in vernacular languages, passes the collected voice data through its machine learning system, and returns appropriate speech and text output. It currently supports 10 Indian languages in over 100 dialects[7].

Dhee AI

Founded in 2017, DheeYantra Research Labs is a start-up that offers conversational chatbots, NLP parsers, transliterations and OCR services for 9 Indic languages. It targets the customer care departments of all vernacular businesses[8].

---

[7] Vernacular AI | Conversational AI Platform. In *Vernacular.ai* website. Retrieved from https://vernacular.ai/

[8] Products – Dhee Yantra Research Labs. In *Dhee.AI* website. Retrieved from https://www.dhee.ai/products/

Agrahyah

Agrahyah provides voice, consulting and content services in and for Indian languages. It is based in Mumbai and most of its members have history in large ICT companies. It has had a wide range of clients, from the French corporation Digene to SBI Life Insurance.

6.5.2 EDUCATION SECTOR

HelloEnglish

HelloEnglish is an Android app that teaches English to the speakers of Indian languages. It provides more than 475 lessons, including exercises for reading, writing, speaking and listening. It also has pronunciation guides. To build vocabulary, it helps its users read the daily news. HelloEnglish currently supports over 50 million learners across the world.

Eupheus Learning

This start-up aims to provide several ed-tech solutions - curricula, online resources, and digital media for different academic skills including language learning. It is used in approximately 22000 private schools across India, by nearly 26 million students.

mGuru

mGuru is a start-up that tries to fix the gaps primary school children have in their understanding of mathematics and English. It has created apps for both subjects; these are available in 7 Indian languages and have reached 1200 users in 6 major cities.

6.5.3 HEALTH SECTOR

Niramai

Founded in 2016, after cancer struck in the families of Geetha Manjunatha and Nidhi Mathur, Niramai provides a computer-based diagnostic tool (SMILE) to check for breast cancer in women. This software essentially relies on a sensing device that scans thermal images for growths and then presents automated reports. Niramai claims that its diagnoses are more accurate than mammography. So far, SMILE has been tested on 4000 patients, and 3 clinical trials have been conducted for publication.

Artelus

This is another start-up that uses image scanning and machine learning to identify health problems. Its main service is the identification of diabetic retinopathy, but it is also developing

tools for tuberculosis, breast cancer and lung cancer. It too has used AI to generate messages and communicate with clients[9].

6.5.4 OTHER

ShareChat

Upon realising that Indian users of most chat applications preferred to type and send other content in their native languages, the founders of ShareChat decided to create a platform specifically for them. ShareChat is a social network and content site that provides all its content in Indian languages, from Marathi to Urdu. It serves around 60 million users per month.

Roposo

Roposo is a video sharing app, much like Snapchat. It is available in English and 9 Indian languages. It started out with a focus on fashion but has since created spaces for other genres of content as well. Its creators aim to have 10 million daily users by the end of 2019.

6.6 SUMMARY

Although still a developing initiative, TDIL has already started the process of commercialising its technology. Automatic speech recognition is being used to help farmers access market information. The cross-lingual information access system allows users to find tourism-related web content in Hindi, English and the language of query. The e-Aksharayan and OHWR tools let users easily scan and recognise printed material. TDIL's TTS software has been put to use in multiple areas, from an Android app to an Indian-language operating system (Indus OS). [10]Recognising the potential of language technologies, several start-ups have entered the ICT market, offering diverse services in Indian languages: AI chatbots, education resources, machine learning software, image-based cancer detectors, social networks, video sharing apps, and more. TDIL must support these start-ups and provide them with strong technical foundations to ensure the development of high software standards. It may also collaborate with international researchers to hasten the state of the art in Indian language technology.

[9] Misal, Disha (2018). 11 Indian Startups Revolutionising the Healthcare Sector with AI. *Analytics India Magazine.* Retrieved from https://www.analyticsindiamag.com/11-indian-startups-revolutionising-the-healthcare-sector-with-ai/

[10] Indus Swipe, Text to speech in Hindi and other languages. In *Indus OS* website. Retrieved from http://www.indusos.com/features/

Chapter 7: Observation and Findings

7.1 OVERVIEW

The chapter delineates the findings of our study on TDIL project in India. The study further has been discussed under the TELOS framework - Technical, Economic, Legal, Organisational and Social parameters. The aim of the chapter is to bring out the consolidated efforts of various collaborating research institutes with the Government of India. A review has been put forward of the already existing tools of TDIL to understand their role and function through different narratives of the five parameters.

7.2 PRIMARY OBSERVATIONS

TDIL, as an initiative, has a lot of potential to pave a way forward towards a more sustainable digital India, which was the idea on which the programme was envisioned. It would also enhance the availability of government services and information by allowing an individual to access information in one's own language, thereby, bridging the communication gap between the citizens and the government. Furthermore, it facilitates the advancement and outreach of various sectors to every region and background of the society. However, even with its increasing popularity and key benefits, it is still yet to make a leap towards its goal of sustainable holistic development.

Firstly, TDIL, as an initiative, is not a solution to all the problems but should rather be observed as a way forward to sustainable development. Secondly, the concept of "language

understanding" needs to be defined well. It is the primary step of Natural Language Processing where understanding a language means knowing the concepts of a word or phrase and knowing how to link those concepts together in a meaningful way. Hence natural-language recognition requires extensive knowledge about the languages and the ability to interpret it. Natural Language Processing in India currently, as compared to the developments internationally, is at in its infancy stage. However, the prospects of NLP in India are far more as compared to other countries. If tools for information processing and communication are available in local languages, it can put an end to "the digital divide" and can pave way to the "digital unite and knowledge for all".

In order to make India a digital India, the concerned authorities, hereby the government, needs to advance onto two major milestones. Firstly, there is an urgent need for special emphasis on research areas. Secondly, development of tools to facilitate smooth functioning and allowing the tools to collaborate with each other to what is called the "plug-n-play" system. The steps are elaborated below:

7.2.1 RESEARCH AREAS

Research areas are a go-to component and a building block of any initiative in the world. Whether, you take up an assignment, project or a task, irrespective of its size and field, one can only advance if the research done is appropriate in order to contextualise its findings and produce applicable knowledge.

Lack of Annotated Corpora

Unlike other countries, in India the basic issue is the multilingualism. Moreover, the same language is pronounced differently in different locations. The same words can also have different meanings at different locations. This makes the researchers work more complex and intricate. In English, and in other languages, many path breaking researches have been done and many pioneering computer-based systems have been developed using language corpora. Importance of language corpora has been recognised by many countries. However, as far as India is concerned, using corpora in language and NLP research is a time-consuming process as it is difficult to capture the fancy of Indian linguistics on account of its diversity. While comparing with British National Corpus (BNC) which contains data obtained from people from all walks of life, we are at a nascent stage. The Speech corpora are also in its primary phase in India. Corpus generation in India is facing several problems due to lack of a centralised authority (a consortium). Many organisations and institutes have collected corpus for their own research activities, but these resources are not available to different groups of people working on corpus generation. This is especially the case for languages which are not widely used thus affecting the availability of resources for them. This has in turn worked in taking the various languages to the brink of extinction.

Lack of NLP Tools

There is an acute scarcity of online lexical resources for Indian languages. Building a Natural Language Processing System without basic lexical resources is almost impossible. This means that anyone trying to build an NLP system has to start from scratch with respect to NLP tools like corpora, lexicons, taggers, dictionaries, morphological generator, POS (Part of Speech) tagger, etc. This is a great challenge for researchers in India. It is a Herculean task which cannot be

achieved by the efforts of a single group. Whatever little data that does exist, has been developed for specific groups and cannot be shared easily. Sharing of resources is the medium that can help NLP projects to take off swiftly.

Lack of Standards

There is an urgent need to popularise standards for the following levels: Script level, Font level, Access level (indexing, sorting, and metadata) and Input level (input/keyboard standards). Moreover, transliteration rules should also to be standardised. Although, some standard drafts, such as the 8-bit ISCII (Indian Script Code for Information Interchange) or 16-bit Unicode for script standardisation, ISFOC (Intelligence Based Script Font Code) for fonts, and INSCRIPT (Indian Script) phonetic keyboard layout have been made and presented, the final standards have not yet been suggested and fixed.

Lack of Evaluation and Certification

Technical evaluation of some of the projects has been taken up by the TDIL and the reports of the evaluation have been submitted for the status update. But a complete assessment of TDIL as a programme is the basic need to see the wholistic picture of language technology development in India. A continuous evaluation with global best practices is required for the programme.

Any technology development leading to formulation of an end user product, should be continuously evaluated and monitored. This should be made available to different groups working on the same issues to take alternative approaches. The best out of these should be taken forward and further developed so that the end product can be ideal. This approach is being adopted in advanced countries to facilitate rapid development in technology. In the case of NLP in India, similar evaluation process is of urgent need, as people working on this field are unaware of the best approaches to be taken and are hence are left confused. An evaluation committee should be set up to identify important approaches and recommend them so that they can be integrated into end-user products. The acceptability of any product is decided by the end users. To prohibit and restrict the use of pirated products, there should be a certification authority for various NLP products. Unfortunately, there is no such certification authority in India.

Lack of Education and Training Institutes

Growth of any technology depends on the proper education and training among people about the same. Once a new technology is introduced, it should be made common and accessible to public, in order to drive the technology forward. Hence there is need for it to be included in the academic curriculum. Since NLP development is a major requirement for e-governance, sincere efforts from government agencies and departments also need to be initiated. In India, NLP based course or training is rare, except for some short-term courses conducted by C-DAC Trivandrum and Indian Institute of Information Technology, Hyderabad (IIIT). But this too is minimal compared to the research centres in foreign universities like Carnegie Mellon University (CMU).

7.2.2 DEVELOPMENT OF TOOLS

Technologies used in the development of the language technology tools become a key deciding factor for the direction the technology will take in future. Since we know NLP in India is still in its

growing stage, the need is to build a high level of expertise of technical resources and information on Indian languages. Development of tools is directly influenced by the development of the corpus. Thus, the available data needs to be converted to digital media and more content needs to be created for the same.

At present, the approach is towards a more sustainable development of technology which is independent of human interaction and interference. Number one, before we further advance with the idea of independent technology, we need to understand the rationale behind it. In order to understand its functioning more effectively, we need to first understand the terminology called "Interoperability". Interoperability, in its literal term, can be understood as the applications that can exchange data and services in a consistent and effective way, facing different hardware and software platforms. In the present world, interoperability is a key success factor. It is the ability to provide interoperable applications and systems which in turn can help in enlarging the market (reaching new potential customers, providing better information, enhancing linking to other sites/info). However, interoperability is not a merely technological issue. We must consider differences in cultures and perceptions of concepts, that is, we have to consider not only a technological interoperability, but a semantic interoperability as well. Thereby, basis is to rely on a consistent framework or a technology which can lead the world of web to its full potential as an international forum.

Other than the above-mentioned observations and key concerns, there is need for us to change our perspective regarding the implementation of these various government initiatives. While the, observations revolve around the need to develop and enhance the tools and technology. We also need to shift our focus towards making it more impactful for the country itself as well. As we have already mentioned about the various ways TDIL can be quite helpful, in regards with the same, there's a lack of commercialisation efforts. Less awareness of the social impact that TDIL shall create and the impact it will have, has also been observed.

7.3 TELOS FRAMEWORK

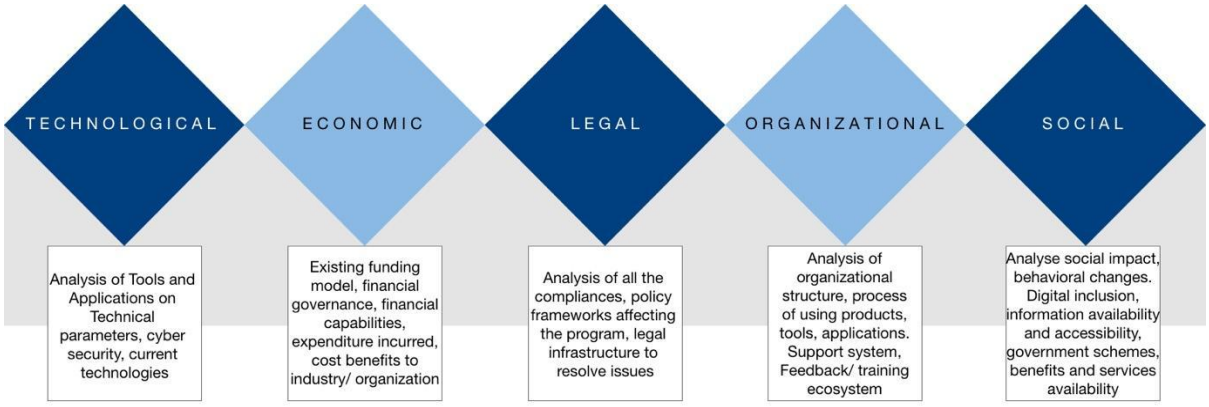


Figure 7. 1: Components of TELOS assessment framework

An In-depth assessment of the TDIL programme encompasses a study based on Technical, Economical, Legal, Operational/Organisational and Social parameters. The parameters along with varied attributes (Fig 7.1) are elaborated as follows:

Technical - The study of technical aspect includes a comprehensive understanding and analysis of Tools and Applications on technical parameters, security, integration, current technologies being utilised in implementation and future scope of emerging technologies

Economic - This study shall include study of existing funding model, financial governance, financial capabilities, expenditure incurred, cost benefits to industry/ organisations and overall impact on economy - if any.

Legal - The Legal aspect of study shall encompass analysis of all the compliances, policy frameworks affecting the programme, legal infrastructure to resolve issues faster, etc.

Operational/Organisational - This study entails analysis of functioning of the ecosystem, organisational structure, and process of using products, tools and applications. Support system, Feedback/escalation mechanisms and training ecosystem for developers, trainers and users.

Social - This study shall analyse impact of TDIL on society including the availability, accessibility and citizen engagement. The study shall also analyse the behavioural changes created because of the usage of TDIL across rural, regional & non-English speaking population. Government schemes, benefits and services availability directly to the beneficiary in its language of choice and communication are some more social impacts that will be covered in the study. Transparency, equal opportunity, removing commute time, empowerment to the citizens in a language understood by them and better governance because of these factors shall also be focussed upon.

7.4 TECHNICAL PARAMETERS

7.4.1 OCR (OPTICAL CHARACTER RECOGNITION)

The Indian OCR system was an outcome of the combined efforts by a consortium of leading information technology institutes in the country that made a simplified robust software with reasonable performance possible. It was not an easy task to achieve as it was riddled with inherent challenges that underlines the development/creation of any language based information technology. One such challenge was that the computing, technology, and training data was not sufficient for it to realise its full potential. The task goals were, hence, set into phases by the concerned consortium.

The very first task within, was to come up with a corpus which was the fundamental requirement for the development, testing and usage of future research in Indian Language document analysis and for the development of OCR. The targets for the first phase were set for a text corpus of 5000 pages for scripts like Bangla, Devanagari, Gurumukhi, Kannada, Tamil, Telugu and Malayalam. The target for character recognition was 98% and for word recognition was set at 90-95%. The processing speed set for a standard PC with 3 GHZ processor with 1 GB RAM was 50 s per 300 dpi A4 size paper with normal text density of 12 point. This target was further modified with an enhanced text corpus of 10,000 pages per script that included more complex images in the above-mentioned scripts. The average character accuracy was hence modified to 95% and the average word accuracy was expected to be 85%.

The technical goals of second phase were:

Automatic script family identification, handling documents with complex layout(Tables, Multicolumn, etc.), processing Multi-colour pages, Multiple fonts (about 25 fonts), italicised and bold font, common symbols and numerals, Bi-lingual (English & Local language) text and higher word level recognition accuracy with appropriate language specific post-processing schemes.

These challenges were handled by dividing the whole project in multiple sub-projects and different groups were given a task as the outcome. The efforts in the OCR were commendable regarding the different kind of tools developed and tested continuously throughout both the phases.

The Indian OCR system was able to improve the script corpus developed in the first phase and make it available for general use through a web interface. They were able to develop additional algorithms for noise cleaning as well as enhanced recognition engines. They were also able to achieve a high accuracy level for the Urdu and Assamese OCR. In addition to this, indexing schemes for the document image were also advanced along with useful HCI for making the system serviceable on different platforms.

7.4.2 TEXT CORPUS

The development of a rich text corpus has always been of high priority for the institutions affiliated with TDIL. Created over several years, the corpus that exists contains information in multiple languages, and pertains to a wide range of domains. Much effort is channelled into organising and making it available to researchers who draw on its linguistic wealth for optimising machine translation, deep learning and other technologies. TDIL has also worked extensively on a speech corpus that is of similar proportions.

As mentioned in an earlier chapter, the usefulness of a corpus depends entirely on the quality, quantity and diversity of language samples contained within it. The Indian government has large data repositories for some languages, spread out over various departments. For example, the Doordarshan archives contains a wealth of news and media information, in text as well as audio format. This data has already been organised through transcripts for different Doordarshan programmes, which match speech to subtitles perfectly. Similarly, the CAG has access to a lot of financial data, the health ministry has health data, and so on. If this data is made accessible to the various TDIL departments, it would greatly expand the size of the existing corpora.

Partnerships between TDIL and different data owner organisations would be a recommended step in this direction. Of course, a concern that must be addressed at all points is that of data security and privacy. If necessary, the information gained must be kept classified, or available only to the computers that use them in machine learning algorithms.

7.4.3 MACHINE TRANSLATION

Machine translation is the language technology that has received the most attention under the TDIL programme. This is because the government realises that the translation of written and spoken content to the various languages spoken in India will empower citizens to expand their intellectual capital and play a more active role in democratic processes. Moreover, it is predicted that machine translation will have a great role to play in the growth of the country's IT sector.

Three major projects have been conceptualised and developed under TDIL's machine translation wing. These have been briefly described in the following table. Each system, along with the assorted tools and resources used during its development, is available for public use on the TDIL website.

Table 7.1Major projects have been conceptualised and developed under TDIL's machine translation wing

Project Name	ANGLAMT	ANUVADAKSH	SAMPARK
Date of launch	2011	2010	2010
Members	Leader: CDAC Noida CDAC Kolkata CDAC Hyderabad CDAC Trivandrum	Leader: CDAC Pune Amrita University Banasthali Vidyapith CDAC Mumbai DDU Nadiad IIT Allahabad IIT Hyderabad IISc Bangalore IIT Bombay Jadavpur University NEHU, Shillong NMU, Jalgaon Utkal University	Leader: IIIT Hyderabad IIT Bombay IIT Kharagpur CDAC Noida University of Hyderabad Jadavpur University Anna University Tamil University IIIT Allahabad IISc Bangalore IIITM-Thiruvanantapuram
Description	AnglaMT is an MT system that uses the rule-based method to translate English text to 8 Indian languages:	Anuvadaksh Phase-I enabled the translation of English text to Hindi, Bengali, Marathi, Urdu, Tamil and Oriya. Phase-II	Sampark is an Indian-Language-to-Indian-Language translator that aims to work with 18 Indian language pairs.

	Assamese, Bangla, Hindi, Malayalam, Nepali, Punjabi, Telugu and Urdu. It was developed as the next phase of an older project, AnglaBharti. AnglaMT has been specifically designed for functionality in the health and tourism sectors.	added support for Gujarati and Bodo. It was initially meant for use in the tourism sector but has since been adapted for the health and agriculture sectors as well.	It currently supports 4 language pairs: Punjabi-Hindi, Hindi-Punjabi, Urdu-Hindi, and Tamil-Telugu.
Test results	Tests were conducted to evaluate comprehensibility (to assess the capability of the engine to understand messages) and fluency (to assess its ability to articulate the same meaning in the target language). It was found that the results were comprehensibility were always higher than those for fluency. It was also noted that the English-Hindi pair produced the best results for comprehensibility as	Tests were conducted using three engines: TAG, SMT and EBMT. All three were used for the tourism domain, while only TAG and SMT were used for the health domain. For TAG, on the basis of comprehensibility, the English-Urdu pair yielded the most accurate translations in the tourism domain, and the English-Gujarati pair yielded the most accurate translations in the health domain. On the basis of fluency, the same results were observed but the resulting percentages were low.	The Telugu-Tamil engine has the highest accuracy (95.43%) for translation. About half of the other language pairs have less than 50 percent accuracy in translations; moreover, when the pairs are reversed, the accuracy of translation changes greatly. For example, Hindi-Sanskrit translation has an accuracy rate of 78.74%, while Sanskrit-Hindi translation has an accuracy rate of 87.7%.

	well as fluency.	For SMT, on the bases of comprehensibility as fluency, the English-Hindi pair was the most accurate translation in both the tourism as well as the health domain.	
End users	1. Various government departments in Kerala, including the State Institute of Languages 2. The Vikaspedia portal (uses English to Telugu module) 3. General public Additionally, the SNLTR Kolkata and WB State Rural Development office have shown interest in the technology.	1. Government Museum, Chennai 2. MAP_IT, Bhopal 3. General public	Sampark has a lot of potential for usage in multiple government departments and other areas. As of now, however, it has not been outsourced.

Each of the machine translation projects listed above works with a sizeable corpus. We are pleased to note that TDIL has recognised the efficiency of the statistical method of translation and is channelling its primary resources in that direction. Of all the language technologies developed, machine translation remains the most significant in the global market, and India has noticed this. Further expansion of the corpus, and an increased reliance on deep learning, should elevate the state of the art of TDIL-led MT technology.

7.4.4 AUTOMATIC SPEECH RECOGNITION

Automatic speech recognition has been researched in depth under TDIL since the 1990s. In 2010, it was used in the Mandi Project[11], which created an application that could accept voice inputs from farmers and provide them with real-time commodity prices and weather forecasts. The project was developed under a consortium of institutes affiliated with TDIL:

- IIT Madras (leader)
- IIT Bombay
- IIT Kharagpur
- International Institute of Information Technology, Hyderabad
- Tata Institute of Fundamental Research, Mumbai
- Centre for Development of Advanced Computing, Kolkata
- Dhirubhai Ambani Institute of Information and Communication Technology, Gujarat
- Birla Institute of Technology, Mesra
- Siddaganga Institute of Technology, Tumkur
- IIT Bhubaneshwar

The application currently works with 9 Indian languages, spread across 12 states: Assam, West Bengal, Uttar Pradesh, Bihar, Jharkhand, Karnataka, Odisha, Maharashtra, Tamil Nadu, Andhra Pradesh, Telangana and Gujarat. It can be accessed by the user by dialling a phone number. This activates a voice recording asking for basic details (desired language, name of farmer's district, name of commodity). Upon receiving the voice inputs from the user, the programme matches the audio with a preassembled speech corpus, then returns an appropriate voice message with the requested details.

During testing, 12 languages were tested. It was noted that the language in which the highest number of calls (700) was made was Gujarati. 89.7% of calls were successful in three attempts. The highest percentage of success (in three attempts) was secured by the Bengali system, whereas the lowest percentage was 58 per cent.

The system is meant exclusively for farmers, although it can be tested by other users as well. Since it provides free service, it has no competitors. It is expected that in future, it will be customised for use in the railways, tourism and health departments as well.

[11] Technology – How it Works (2018). In *Mandi Project* website. Retrieved from https://asrmandi.wixsite.com/asrmandi/technology

7.4.5 CROSS-LANGUAGE INFORMATION ACCESS

In recent decades, much information in India's vernacular languages has been uploaded onto the internet. This includes articles, e-books, reports, works of fiction, audio/video content and so on, all of which can be accessed with considerable ease by the native speakers of each language. However, the access to information in one vernacular language by the speakers of another vernacular language is not a simple process. For the average Hindi or Bengali speaker, it is very difficult to even locate online material in Malayalam or Konkani, let alone understand it. TDIL's Cross-Lingual Information Access system (CLIA), Sandhan, aims to remedy this.

Sandhan was launched in 2006 by a consortium of 12 institutes - IIT Bombay (leader), AUKBC Chennai, AUCEG, Chennai, CDAC Pune, CDAC Noida, DAIICT Gandhinagar, Gauhati University, IIIT Bhubaneswar, IIIT Hyderabad, IIT Kharagpur, ISI Kolkata and Jadavpur University.

Sandhan is available for public use on the TDIL website and as an Android application. It allows users to search for key phrases in 9 languages – Bengali, Hindi, Marathi, Punjabi, Tamil, Telugu, Assamese, Oriya and Gujarati – and returns relevant online documents in Hindi, English and the language of query.

It essentially relies on machine translation and web crawling technology to look for results in different languages. Although the framework for the system has been firmly established, much work remains to be done in order to widen the scope of documents accessible via Sandhan's search engine.

7.4.6 TEXT TO SPEECH

The text to speech software for various Indian languages was initiated with the aim to develop varied monolingual and bilingual text to speech synthesis system in them and make them available to the various consortium members and public. This project began in 2009 through a consortium of leading technical institutes of the country under the leadership of Prof. Hema Murthy of IIT Madras.

The project, till date, has been initiated and developed in two phases. Phase I of the project used Festival-based speech synthesis for Bengali, Hindi, Tamil, Telugu, Malayalam and Marathi. Phase II of the project commenced in 2012 employing HTS based statistical speech synthesis for 18 Indian languages.[12] These include Hindi, Bengali, Marathi, Tamil, Telugu, Malayalam, Gujarati,

[12] https://goinggnu.wordpress.com/2017/03/07/how-to-ask-iitm-to-release-indictts-as-freeopen-source-software/

Oriya, Assamese, Rajasthani, Manipuri, Kannada, Bodo, Konkani, Punjabi, Kashmiri and Urdu. The software also has a 40-hour database with text transcription for one male and one female speaker. The best performance was shown by the male speech for Tamil which received 3.95 MOS against a target of 4 while the poorest performance was shown by female speech synthesis in Bengali at 2.81 against a target of 4. All these were calculated on a scale of 0 to 5.

They have also developed android applications in 4 languages (Hindi, Tamil, Telugu, and Indic) to make the TTS service available on android platforms. The TTS has also been integrated with Non-Visual Desktop Access (NVDA) screen reader and Optical Character Recognition System (OCRA) to enable visually challenged users to interpret and perform computer operations with audio interface. It can be integrated with the following screen readers:

1. Windows 7, Windows XP based NVDA (2010.1) Screen Reader for Indian Languages

2. Linus ORCA: Screen Reader for Indian Languages

TTS works for both these operating systems with a voice quality with mean Opinion Score between 3.7 to 4, the highest at 4 being for Tamil, Indian English, Hindi, Marathi, Bengali, Malayalam and lowest at 3.7 for Kashmiri, Bodo, Manipuri, Urdu, Punjabi, Konkani. It can also be used as a Web based service available on Mozilla Firefox plug-in and Google Chrome plug-in allowing it to give the power of speech to browser.

The TTS software, however, is available as open source only to the consortium researchers and academicians and not the public at large. The android applications released are at a preliminary stage and without proper license and attributions to base, open source HMM HTS systems.[13] The web version is nevertheless efficient with human like voices. There is a need for this software to be made available for the people at large so that it creates an ecosystem for further innovation in the field and not make people to start from scratch.

7.5 ECONOMICAL PARAMETERS

7.5.1 INTRODUCTION

In this world of digitisation and the internet, it is hard to imagine our lives without Information Communication Technology (ICT) and the different ways in which it has made our lives more convenient and easier. ICT today has seeped through all facets of the global economy and society. It is transforming how all goods and services – not just digital or information goods -are

[13] ibid

designed, produced, distributed and sold, so much so that to be part of the global economy, one has to be a digital citizen. India, as a nation, has been known for its contribution to Information Technology and had made a name for itself on the global platform. But this technology and its benefits have, however, remained untouched by a large section of our own population. This has been majorly so because most of these developments are created using English as the language of choice and there is a need to build them in other languages given the multilingual nature of our country.

India is currently in the phase of expanding its reach in the field of Information Technology by making it user friendly and accessible among the different linguistic groups within the country. This technology has the potential to play a tremendous role in fulfilling the aspirations of millions in the country if it is based on the ideals of "broad-based" economic growth that provides more effective public services quickly. This change can be a major mover for economic growth in the country. India is emerging as the second largest digital market after China, with over 12% of the internet users in the world, majority of them being non-English users whose numbers are expected to rise in the years to come. Localisation in this field can, thus, be a profitable venture with an ever-expanding demand and market. The various initiatives undertaken by the TDIL programme have been worked upon over the years, providing various services, applications, and tools to facilitate this process of localisation.

7.5.2 BUSINESS MODEL FOR INDIAN LANGUAGE TECHNOLOGIES

A business model is a plan for a successful operation of a business that identifies its sources of revenue, the intended customer base, products, as well as the details of financing. This becomes especially necessary in the sphere of ICT development within a country as innovation within this sector can be the primary mover for the future of economic growth. It also plays an important role in creating and maintaining new jobs. Thus, a flexible business model for ICT based products, services, and application is fundamental so as to not hinder innovation and creation of an ecosystem that promotes the same. The TDIL programme under the Ministry of Electronics and Information Technology (MeitY) should also allow for a flexible business model where the various services made available by the programme are provided at no or minimal cost. This would encourage and allow a space for innovation within the ongoing ICT programmes, subsequently contributing to the growth of the country's economy.

7.5.3 IMPACT ON RURAL ECONOMY

The localisation of ICT in its present form can have a huge impact on the rural economy of our country. The various language information technology tools and applications under the TDIL programme have an immense potential to raise their farmers' income and productivity. Localisation can give the farmers access to all the latest developments in agrarian sector like hybrid and genetically modified crops, precision farming that uses sensors and GIS based soil, weather and water data that can steer their farming decisions as well as mobile internet based farm extension and market and information services. These alone can help create more than $45 billion to $80 billion per year in additional value that this sector could realise by 2025, thereby making the goal of achieving $5 trillion economy by 2024 a possibility.

Availability of these services in a language that one is comfortable with, will eventually help farmers reach their full potential. It will also give them access to the various policies and schemes that the government has brought out for their assistance and relief. It would further help them bring about improvements in the storage and distribution systems that could subsequently reduce the losses they occur once the crop has been harvested. This can save as much as $32 billion per year in 2025. These are especially important in the context of the agrarian crisis that country is currently going through and given that is still employs the majority population of the country. All of these developments can have a huge impact not only on the farmer's income, but also contribute immensely to the economic growth and the GDP of the country.

7.5.4 IMPACT ON THE REGIONAL ECONOMY

The development and localisation of language technology can play the role of empowering the economy of a region as well. These technologies work to provide the necessary basis that can uplift a region that had otherwise been isolated because of the language barrier that subsequently hindered communication with the region. These initiatives can help the people in these regions access opportunities, amenities and resources which they are currently not able to, due to the language barrier. The advancement of Language Information Technology can further help promote the local businesses in the region and create a large market for them which can bring them more profit (empower the region) and eventually contribute to national economic growth.

7.5.5 EMERGENCE OF NEW BUSINESS AREAS IN VERNACULAR LANGUAGES

The process of localisation of ICT can also contribute to economic growth by creating newer business areas and job opportunities in the vernacular language sector once it is developed and working. The rapid spread of these technologies has already created a favourable environment for the development and creation of newer business opportunities and innovation in the sector which were earlier not possible. The proliferation of these into various Indian language interfaces can further this process, where different aspects of the same would require developing a system that can help facilitate the development of these technologies into a particular language.

This can only be made possible by providing the data in various languages for services that the TDIL programme currently offers. These include translation technology services (comprising of input and output of data as audio, images, text, and documents), automatic digitisation, chat bots for support (especially in the customer care sector), as well as providing information and grievances for different applications, products and services in various sectors. All of these would require data in various languages from the people who know them. The development of the ICT in a diversely lingual country like India can thus prove to be a bane for its economy and future economic growth.

7.5.6 SECTORIAL ECONOMIC IMPACT

The localisation of ICT would also have a powerful effect on various other sectors of the Indian economy like education, healthcare, e-governance, citizen engagement, financial services/technology, as well as e-commerce. All these sectors of the country's economy account for about 45% of the GDP and employ over 60% of the population. The development of Language Information Technology in these sectors has a great potential in the future for India's economic growth.

- Education and Skills: the development ICT into local languages can help improve the quality of teaching and also help remove the disparity in the level of knowledge that exists due to the lack of access. School performance can be improved through e-administration, digital identity-based attendance systems, and online teacher certification and training. It can also be used to improve learning by facilitating a more hands-on approach at not just the school level but also at the level of higher education. This will eventually help improve the productivity of future workers, and this higher

productivity of more skilled workers will contribute to the economic growth of the country by generating over $40-$90 billion per year by 2025.

- Healthcare: India today, in proportion to its population, has about half the doctors, nurses and healthcare centres in the country. The development of language information technology in the healthcare sector would help extend the care to even the remotest of the areas in the country. It would also help all the people in the country have equal access to modern amenities and developments in the field of healthcare. The total value of empowering technologies in health care could be $25 billion to $65 billion per year by 2025.
- e- Governance: e- governance is a tool of information communication technology that works to provide access to different government services, information exchange, communication transactions, as well as to bring together various systems and services between the government and citizen, and the government and business. The vernacularisation of these services into various Indian languages would help people even from the remotest part of the country to access these and subsequently become able to contribute to the betterment and economic growth of the country.
- Financial Technology: the Indian banking sector already uses technology to digitise its operations and services to its customers such as online brokerage, mobile banking, online insurance sales, etc. This sector however still faces the limitation of financial inclusion within the country as just over a quarter of the country's population has access to a bank account. Technology in this sector can help facilitate efficiency and the process of localisation of this technology would bring the advantages of these services and tools to the masses at large. This would also save the government around $100 billion per year it spends on paper-based plans and could spell out as economic growth of over $32-$140 billion per year by 2025.

7.6 LEGAL PARAMETERS

In order to develop a deep understanding of the legal component of the TDIL programme, we need to first understand the present legal and constitutional structure of Indian legal system related to Indian languages and the issues arising out of their implementation in various sectors of Indian economy, both at the local as well as the global level. There is an immense demand for building up an ecosystem of compliance encompassing the national and international paradigms

in order to meet the near future challenges. With the advancements of internet, the world is no more global, it's becoming rather 'Glocal'. We need to interlink the legal parameters with the local and global economic, social, technological, organisational and political models. None of them can be kept in a water-tight chamber to get hold of a holistic view that alone will be able to address the issues and challenges of present as well as of future.

7.6.1 OVERVIEW

In the legal component, we start with discussing the legal provisions of Indian languages by highlighting the present Constitutional Acts and Policies related to Indian languages. We further go on to discuss the conflicts arising out of the legal compliances related to e-commerce, contract management, government policies, e-medicine, e-farming, agro-markets, fin-tech services, e-governance models and many more. Issues related to national and global standards are also highlighted to develop a clear understanding for developing a robust ecosystem of legal acts, policies and compliances.

Legal Provisions of Indian Languages in India

Constitutional Provisions

7.6.2 COMMITTEE ON OFFICIAL LANGUAGE

The Official Languages Act, 1963 declared Hindi in Devanagari script as the Official Language of the Union under Article 343(1) of the Indian Constitution.

It was envisaged that English will continue to be used for executive, judicial and legal purposes for an initial period of 15 years i.e. till 1965. The period of 15 years was prescribed after a detailed deliberation so that necessary arrangements could be made for smooth language transition.

The Committee has so far submitted reports in Eight parts, but in the very first part of the report itself, it decided to go into the translation arrangements and various aspects thereof in the offices of the Central Government. This highlights the very importance of 'Translation' even in the year as early as 1987.

Indian languages have the potential of being used as an economic, religious and political communication link. Thereby developing a need for an effective translation mechanism catering to the diversified scripts and dialects. At micro-cosmic level, there are numerous regional languages with millions of users waiting to be connected through knowledge sharing platforms

and there stands a huge market opportunity as well due to the lack of effective translation systems and well-developed text corpus.

7.6.3 FUNDAMENTAL RIGHTS

Under Article 29(Cultural and Educational Rights), Protection of interests of minorities- Any section of the citizens residing in the territory of India or any part thereof, having a distinct language, script or culture of its own shall have the right to conserve the same.

Under Article 30, Right of minorities to establish and administer educational institutions: All minorities, whether based on religion or language, shall have the right to establish and administer educational institutions of their choice.

As the State recognises linguistic minority, it is the duty of the State to provide full support for the development of the language of any region to its full potential so that all the possible opportunities can reach to the grassroots without any discrimination.

7.6.4 FUNDAMENTAL DUTIES

To promote harmony and the spirit of common brotherhood amongst all the people of India transcending religious, linguistic and regional or sectional diversities; to renounce practices derogatory to the dignity of women.

Thereby the duty of each and every individual is to make our society a well-connected web by promoting and contributing to the research and development of a sound translation systems for bridging the gap created by linguistic barriers.

http://www.languageinindia.com/april2002/constitutionofindia.html

7.6.5 RIGHT TO INFORMATION ACT

Information empowers and enables people, pushes them towards exercising their legal, social, economic and political rights. Almost every society has recognised the same by way of putting in place the mechanisms for free flow of information and ideas so that people can access them whenever it is required without too many procedures. Information accessibility will lead to the dawn of a new era in our processes of governance, an era of performance and efficiency, benefits of growth will flow to all sections of the society, developments in the field of education and science & technology will reach the most humble and downtrodden, eliminate the scourge

of corruption, and will bring the common man's concern to the heart of all processes of governance and fulfil the hopes of the founding fathers of our Republic.

The Right to information (RTI Act 2005) is a legal right for every citizen of India. Under the provisions of the Act, any citizen of India may request information from a "public authority" within thirty days. The act also requires every public authority to computerise their records so that the citizens need minimum recourse to request for information formally.

But the process of seeking information is not that easy. As per RTI Act, we can seek information in English, Hindi or Local Language. So in Maharashtra, one can seek information in Marathi, besides English or Hindi. Unfortunately, reverse is not true, i.e., departments/public authorities of state governments are not providing information in these 3 languages. It has been observed that some the departments/public authorities of state governments are providing RTI information only in local language through their official website.

There are so many examples from state government official website. Unfortunately, if one doesn't understand local language of that particular State, then RTI Act stands less efficient. This be eased by the use of language technology to some extent.

A government "of the people, by the people and for the people" is the spirit and essence of a democracy and its accountability towards people. Even the World Bank document of 1992 on Governance and Development in its quest for 'good governance', identifies accountability as well as transparency and information to constitute as one of the most specific aspects of 'governance'. The visibility of deprived communities increases on the political map, and their interests can be realised by giving them the power to seek information. This goal can be realised if the RTI Act takes full account of the various translation-based constraints that defeats the whole purpose behind 'Information for all'. With sound translation systems and fully developed text corpora available in all regional languages, it will encourage the mechanisms for free flow of information and ideas for a truly empowered society.

7.6.6 NATIONAL DATA SHARING AND ACCESSIBILITY POLICY – GOVERNMENT OF INDIA

A large quantum of data available either in digital or analogue forms generated using public funds by various organisations and institutions in the country remains inaccessible to the public, although most of such data may be non-sensitive in nature and could be used by public for scientific, economic and developmental purposes. There has been an increasing demand by the community, that such data collected with the deployment of public funds should be made more

readily available to all, for enabling rational debate, better decision making and use in meeting civil society needs. The NDSAP policy is designed to promote data sharing and enable access to Government of India owned data for national planning, development and awareness.

The motivation behind this policy is the United Nations Rio Declaration on Environment and Development, Principle 10: "Environmental issues are best handled with the participation of all concerned citizens, at the relevant level. At the national level, each individual shall have appropriate access to information concerning the environment that is held by public authorities, and the opportunity to participate in decision-making processes. States shall facilitate and encourage public awareness and participation by making information widely available."

Thereby, an Open Government Data "OGD Platform India" is a platform for supporting open data initiative of the Government of India which was created on 17th March 2012. The portal is intended to be used by the Government of India ministries, departments and their organisations to publish datasets, documents, services, tools and applications collected by them for public use. The base "Open Government Data Platform India" is a joint initiative of Government of India and Federal government of the United States.

The current version, launched on 11th December 2014, is the stable release of the platform. After the launch of the Digital India programme in 2015, "Open Government Data Platform India" has been included as one of the important initiatives under Pillar 6 – "Information for All".

NDSAP 2012 is already taking care of some of the legal aspects of language technology domain by providing the right sharing and accessibility policies. The next version of NDSAP can have comprehensive policies and directions that can nurture the legal needs that will be created at the time of the language technology tools and resources implementation in the day to day use of the society for the ground level empowerment.

7.6.7 NATIONAL MISSION ON NATURAL LANGUAGE TRANSLATION

The Ministry of Electronics and IT will soon place before the Union Cabinet, a Rs. 450 crore proposal for Natural Language Translation. It is one of the key missions identified by the Prime Minister's Science, Technology and Innovation Advisory Council (PM-STIAC).

The mission aims to make science and technology accessible to all by facilitating access to teaching and researching material bilingually in English and in one's native Indian language. To

achieve this, the government plans to leverage a combination of machine translation and human translation. The mission would also help students, teachers, authors, publishers, translation software developers and general readers.

To overcome the language barrier for the citizen, the government has planned to set up a language technology environment which will involve the Central agencies, State agencies and start-ups in language technology. The start-ups can help expedite the work to build implementable solutions to help provide services. Further, the translation activities can also help generate employment for educated unemployed.

7.6.8 VARIOUS CONFLICTS IN THE WORLD OF WEB

In a fast-moving world of ideas where the most complex tasks are performed over just a click, we need to be better equipped with the challenges that stands in the way of achieving such a convenient technology. With the global brands and services penetrating the most remote areas, there is a need to develop better preparedness for safeguarding the interests of the poor and uneducated people who are unaware of the legal nuances in case of customer rights violations.

Observations

There is a need for a strong legally compliant system for various issues arising out of e-commerce compliance, online contract management, government policies, e-medicine, e-farming, fin-tech services, government policy distribution, various e-governance models, and many more. These nuances can be better understood by the help of an example. There can be a case that, while availing an online Fin-tech service, due to a minor translation errors, choosing one stock might get misinterpret leading to stock holding of some other stock, leading to losses. In such a scenario do we have a proper data source to re-confirm and verify at the time of complaints. Such legal and technical provisions need to be better addressed by providing a legal framework that can solve such kind of scenarios effectively.

TDIL & its research teams have done good work by building language standards and collaborating with Indian and Global Standard agencies. They are able to put up a solid foundation for language standards and compliance platforms, but we stand slightly away from finalising the standards for a risk-free environment of compliances.

With well-developed standards incorporated with our existing Acts and policies, a robust legal framework can be envisaged to tackle the future challenges coming out of various legal constraints in the world of Web providing various services in a diversified country like India.

7.7 SOCIAL PARAMETERS

7.7.1 AVAILABILITY

TDIL aims to build a network that is readily available to every citizen in their local language. The vision is to make digital initiatives available to every citizen in the country. Almost all ICT applications are available in English language only. In multilingual countries like India, one of the major set-backs in availability of ICT services is in terms of language. In order to make the information readily available, the TDIL initiative is proven to be a step ahead, into bridging this particular gap. Not only for the ease of usability, but also to empower the citizens and make them more aware, it is necessary for the information to be readily available in a language that one understands. Important information with regard to government services being available - e-banking or net banking services, healthcare services, agricultural feedbacks and services, etc. - are being provided. However due to lack of availability of content in local languages, it hinders with the productivity of the same. To disseminate all this information to all the citizens, it is important for the information-processing to be available for use. Manual translation is slow and expensive, so catering to the needs of people with non-Hindi or non-English vernacular becomes difficult for the government, and therefore it becomes pertinent to develop efficient technology.

7.7.2 ACCESSIBILITY

While we are providing tools to access the information, our goal of developing technologies of machine translation shall also aid us in making information accessible. The access to information is the primary goal of developing language technologies for Indian languages. After the technology is made available, it is important to make it accessible by providing it free of cost so that people do not hesitate in using it. Accessibility has to be ensured by making the technology available in a medium which people have available to them easily. This empowers a citizen by giving them power to get information in a language of one's own preference and cannot be duped or taken advantage of by blocking the way to information. With the help of technologies like Angla-Bharti translation or Sampark engine, one can translate across English-to-Indian

languages and Indian-to-Indian languages; these engines have been specifically fine-tuned with health and tourism domain to produce accurate results. Such technologies will make it easy for the citizens to have access to information when they move to another state or region for employment, health or tourism purposes. Any citizen should be able to access information that affects them or can help them from anywhere and at any time. This way the accessibility of tools and information will play a huge role in empowering the citizen, as well.

7.7.3 TRANSPARENCY

Transparency is identified as the key or the most important factor to determine the parameters of governance. Today, the entire world is aiming towards the goal of achieving a more transparent and accountable government or organisation, which in return takes care of the interests and identities of the beneficiaries of that information. One tends to display information in public to gain confidence of the concerned recipients of the information. If the data is displayed in a language one does not understand then it upsets the idea of it being transparent, since it has not been able to deliver itself. Transparency is a crucial element for imbuing harmony between the State and its people - that a government conducts its decisions and actions in a transparent manner is one of the key ideas of a democratic culture. Similarly, to conduct business between two parties, transparency is crucial to strengthen the decision-making capacity of an individual. Therefore, if there is a relevant technology that can bridge the gap and serve the purpose of transparency, then the individual becomes empowered and takes the information more seriously.

7.7.4 EQUAL OPPORTUNITY

If the information is made available and accessible in local languages, this not only ensures a way forward towards active participation but also gives the citizens an equal or same opportunity to make use of that information. Hence, language would no longer be a hurdle to a person coming from a less informed background. With properly developed tools for digital access of all the distinct languages in the country, no one will be left out from the benefits of information technology. A culture which is able to offer transparent information will pull people from all sections of society and initiate them into the world wherein each individual can thrive after being given equal opportunity. With governments using online portals for tenders and projects, classroom education adapting to the growing relevance of e-learning, career opportunities being widely publicised on online platforms, and most importantly, with more and more people trusting internet to connect them to the world, it becomes important that one can

avail the information in a language of their preference. TDIL has developed tools like Uni-coded Fonts and distributed Language CDs which has given people an opportunity to use computer and input data in the script of their language. Such initiatives empower people to store data and share it in their own language. This shall bring forth equal opportunity for all the sections of society as all the means of empowerment - education, employment, business, health - will be uniformly present for everyone without any caste, class, regional, religious barriers. The dissemination of knowledge in different languages and the ability to procure information by using machine translation software like Anuvadaksh developed by TDIL, will empower the citizen to explore and realise his potential.

7.7.5 TRUST

Any data or information which is not available in one's preferred language is always accompanied by a general mistrust. This in turn makes one mistrustful of another individual or agency whose language he or she is incapable of comprehending, which will surely become a decisive hurdle in any business being conducted between two parties. The linguistic difference brings in a feeling alienation from one another which can be bridged only by doing away with the rift and making information seamless. These problems impede the free flow of information and disrupt communication. Therefore, when a reliable technology like Angla-Bharti or Anuvadaksh is in place, which can translate the information given in a language beyond our understanding, it develops trust and strengthens communication. The trust factor is important because one party could be more vulnerable to being duped or taken advantage of by the other, especially where social or economic hierarchies are concerned. If the tools for language technologies developed by a government initiative like TDIL are widely available and used by public, they become a trusted resource by shedding off the burden of mistrust from the other language. With the presence of such tools, communication and information reliability across languages will increase and people shall engage more confidently in commercial and social exchanges with each other, which will eventually develop a harmony in communication among people. Moreover, trust building shall also enhance the transparency, thus, building a stronger relationship between two or multiple parties.

7.7.6 UNITY

India has always fostered upon its "unity in diversity" since the beginning as an independent country. Thus, a multilingual subcontinent like India needs to inculcate harmony between the speakers of different languages. This becomes difficult if they are deprived of the benefits of

digitalisation because there are not enough tools for them to access the internet, smart-phone or the computer in their own language. The proposed unity can foster only when the users of all the official languages have the same level of technologies and tools developed for their languages. This sense of unity will help in availing benefits not only from one's region, but also from the other parts of the country which bring about a more vibrant and wider space for engaging in economic and cultural activities with each other. A search engine like Sandhan gives search results in multiple Indian languages, which will help a user search for relevant data from any other available language even with the help of a keyword.

7.7.7 RURAL/REGIONAL/NON-ENGLISH INCLUSIVENESS

When we talk about development of ICT, it isn't limited to just the urban population but to the rural as well. Today, the rural participation is actively increasing and so is their engagement on various online portals. With the implementation of TDIL programme, the same shall ensure inclusiveness of people coming from various backgrounds. Availability of information in local language can further enhance this participation from the rural areas. It can be helpful for growing e-commerce, wherein, people can sale-buy and purchase products and services without having any language barrier. Not only will it attract the international market, but services like Amazon and Flipkart shall also be an active part of it in the future. Moreover, this shall pave a way forward for giving people from rural background a chance to take their local business to the national and international markets, further facilitating the initiative of "Smart Villages". Healthcare facilities, especially for rural areas, agricultural policies and various services information would be readily available. Therefore, this shall also help stem the migration of rural citizens to urban centres.

In a vast country like India, where various regions are left out or neglected because of geographical and linguistic barriers - like North East states, Kashmir Valley and Southern States - shall also find their voice and feel more included as a part of the nation which provides services in their said local languages.

Moreover, it is to be also noticed, when we talk about inclusiveness, the same could be focused in terms of English and Non-English speakers. It is not necessary that the urban population is very well versed with English, howsoever, with TDIL in action this shall also remove the barrier of language and a person sitting in an urban setting is also able to access government services in her/his own local language.

7.7.8 CITIZEN ENGAGEMENT

Citizen engagement or public participation, in literal terms, can be defined as participation of the public in policy making with the government. Citizens play a critical role in advocating and helping public institutions become more transparent, accountable and effective. Today the world has been advancing towards a more citizen-centric approach, wherein the solutions are not only effective and innovative from the change-maker's perspective, but also a contribution from the people who are directly affected by these policies. In democratic countries like India, wherein the basis of governance is built on "of the people, by the people and for the people", engagement of citizen is a tremendously crucial part. With the implementation of TDIL, the same can be enhanced and facilitated in a better way. Cities like Bengaluru and Hyderabad have developed various technologies which focuses on citizen engagement for redressal of public grievances, providing citizen services, such as applying for Driving License or Passport directly from your smartphone. Such initiatives could be further enhanced if the same applications are available in one's local language. Citizens then can also give their feedbacks and comments easily, therefore, empowering the citizen as an active participatory agent in the policy-making process.

7.7.9 GOOD GOVERNANCE

Good governance has 8 major characteristics. It is participatory, consensus oriented, accountable, transparent, responsive, effective, efficient, equitable, inclusive, and follows the rule of law.

As mentioned above, we have largely covered the major implications of TDIL in the model of Good Governance. The model of good governance in any nation ensures minimised corruption, and inclusion of minorities as well. When we break the barrier of languages and make services and information available and accessible to every person and in every corner of the country, the government has already established a strong foothold towards the way of developing a framework of good governance.

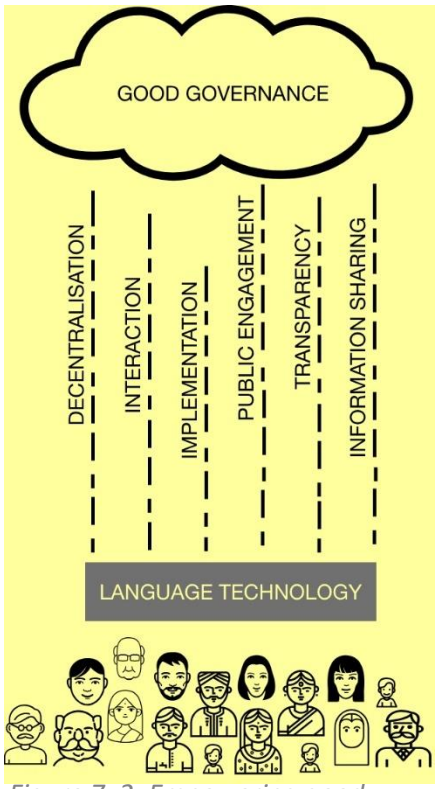


Figure 7. 2: Empowering good governance parameters using LT

7.8 SUMMARY

The TDIL programme was evaluated through the TELOS framework, which includes Technical, Economic, Legal, Operational/Organisational and Social parameters. The report team first considered each TDIL initiative on technical grounds, subdividing it into subprojects. It judged not only the current state of each subproject, but also the progress it had made since its conception, making notes on the areas that needed more focus in future.

Next, the team assessed the business model of TDIL, keeping in mind its impact on rural and urban demographics, rising businesses, and various sectors. It was recommended that the developed language technologies be made available to the public at minimal or no cost. The team calculated that language technologies, if properly utilised, could increase the size of the Indian economy by at least $150 billion per year. This would make it possible for the country to become a $5 trillion economy by 2024.

The need for multilingualism has been a recurring theme in the history of Indian law. Early on, the Committee on the Official Language recognised the importance of translation. Under Articles 29 and 30 of the Constitution, the State is responsible for aiding the development of regional languages in all spheres. In order for democracy to exist, information must be accessible to speakers of all languages in India. It is in this interest that the government adopted the National Data Sharing and Accessibility Policy, and later the Digital India programme. The Ministry of Electronics and Information Technology now intends to put forward a Rs. 450 crore proposal for machine translation projects, under the new National Mission on Natural Language Translation.

To conclude its assessment, the team highlighted the social advantages of the TDIL programme:

- It makes digital information available to all citizens.
- This information is accessible, as it is delivered in the languages of the people.
- The programme also builds a culture of transparency, which increases faith and trust in the government.
- It provides more opportunities for upward social mobilisation.
- It strengthens national unity.
- It includes local identities in the newly digitalised world.
- It allows citizens to directly engage with democratic processes.
- It leads to good governance.

Chapter 8: ANALYSIS OF SURVEY

8.1 CHAPTER OVERVIEW

This chapter aims to cover the analysis of the primary research of TDIL gleaned through survey technique in stage III, phase-II of the study.

The vision of the survey was to understand from the project team's perceptive, as to how they have been impacted by the TELOS based attributes of the system, namely, Technical, Economic, Legal, Organisational/Operational and Social. These have been analysed through a strenuous data collection tool. Legal aspect has been kept out of the survey tool as the study of the legal aspect has more to do with the understanding of the legal framework and compliances that can address the future legal needs of the language technology. This study is comprehended in the last chapter in legal findings.

The study further explains the various aspects of TDIL included in each of the five pillars and shows how each of these have created an impact to strengthen the language technology environment. The output of 'Stakeholders' consultations, undertaken through detailed meetings and round table interviews have also been summed up in the chapter.

8.2 INTRODUCTION

The existing structure of the project from the management perspective can be understood as resting on five pillars as illustrated earlier in the assessment framework. Each of the pillars cover

some key components related to research areas, deployment, commercialisation efforts, legal compliances, acts and policies on LT, impact on social and economic sector through applicability in multiple areas across stakeholders and lastly, managing the resources to make it a lucrative initiative.

It is important, at the outset of this chapter, to understand the methodology adopted and process undertaken. The methodology includes

1. Semi-structured interviews with the management
2. Information received from TDIL office records
3. Various reports and vision documents of TDIL
4. Secondary research of related websites
5. Structured survey questionnaire conducted for the research institutions (refer Annexure).

It is important to understand that the study cannot be validated without the affirmation from the most important stakeholder which are, in this case, knowledge institutions doing the research on TDIL projects. The core facts of the study have been taken from the feedback of project teams of TDIL. Of the institutes to whom the survey was sent to, **14 responded to the same, which is a base sample for the impact assessment analysis.** As a result, this study is based on the data collected and collated from these research teams. This study has been based on the multiplicity of methods which included survey tools designed primarily on the conceptual framework. **Through open ended and close ended questions, the survey focused on capturing the impact of TDIL on five pillars of the study.** The information thus obtained helped to understand the management perspective and assess whether the vision of TDIL has been achieved or not. It is of crucial importance to highlight that in the design of the survey tools, a variety of questions were asked while adhering to the pillars of the study. This was done with the intention to seek responses for each of the pillars while studying the possibility of having a holistic perspective emanating from various vantage points.

The findings of the study have been presented in the subsequent sections of this chapter. These have been presented in the order of Technical, Economic, Organisational/Operational and Social which also form the four pillars out of five pillars of the study.

8.3 TECHNICAL

In the course of the study, Technical components of TDIL has been ascertained through twelve important consideration which are as follows:

1. Usability
2. Efficiency
3. Reliability
4. Portability
5. Innovation
6. Ease of Integration
7. Programming Skills
8. Emerging Technologies
9. Supported Devices

10. Comparison with Other Technologies
11. Gap in Facilities for the Development
12. Future Thrust Areas in LT

8.3.1 USABILITY

As mentioned above, TDIL has a major impact on the society today and in future as well. Thus, it is essential to assess the projects performance based on the usability.

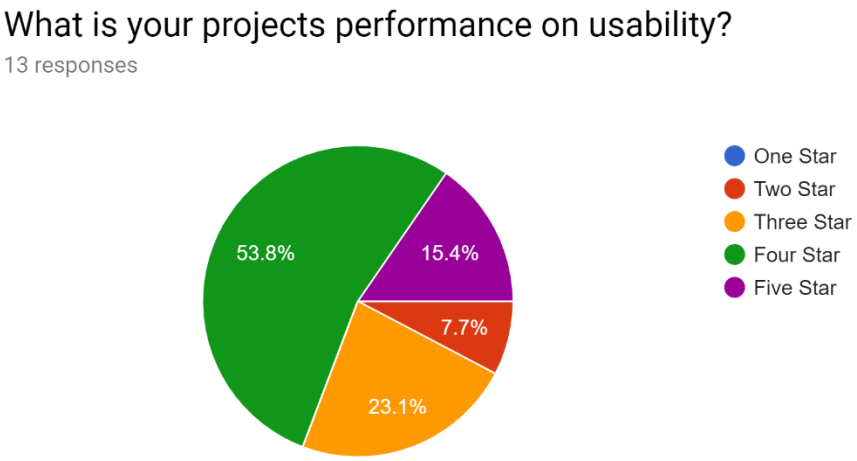


Figure 8. 1: Project performance on usability pie chart

On a scale of 1-5 (wherein 5 being the most and the 1 being the least) almost 16% rated their projects performance a 5-star rating. More than half (almost 54%) of the developers find their projects to be efficient enough with a rating of 4 stars. Rest 23% found it to be average, i.e., three-star rating and remaining 6 -7% rated it below average.

8.3.2 EFFICIENCY

On a scale of 1-5 the efficiency of the projects was measured, wherein 5 being the most efficient and 1 being the least efficient.

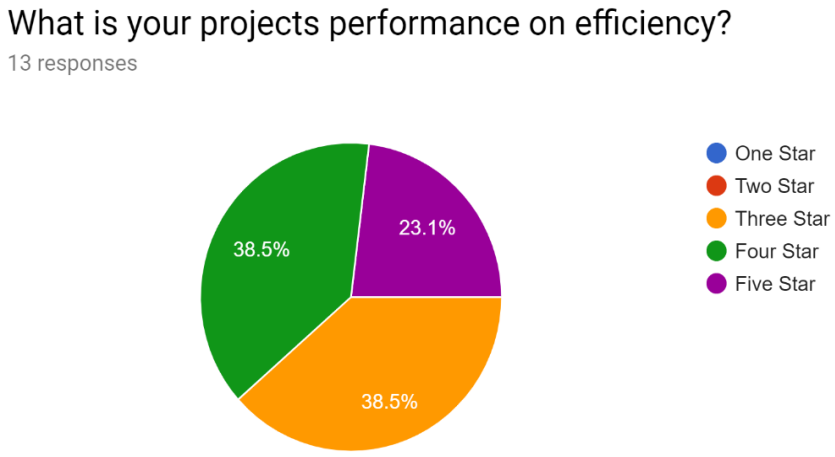


Figure 8. 2: Project Performance on efficiency pie chart

According to the responses received, more than 20% find the projects to be most efficient with a rating of 5 stars and the rest 80% were equally divided in rating 3 and 4 stars respectively.

8.3.3 RELIABILITY

Similarly, on a scale of 1-5, with 5 being the most and 1 being the least rating, the question was raised on mapping out the reliability of the projects.

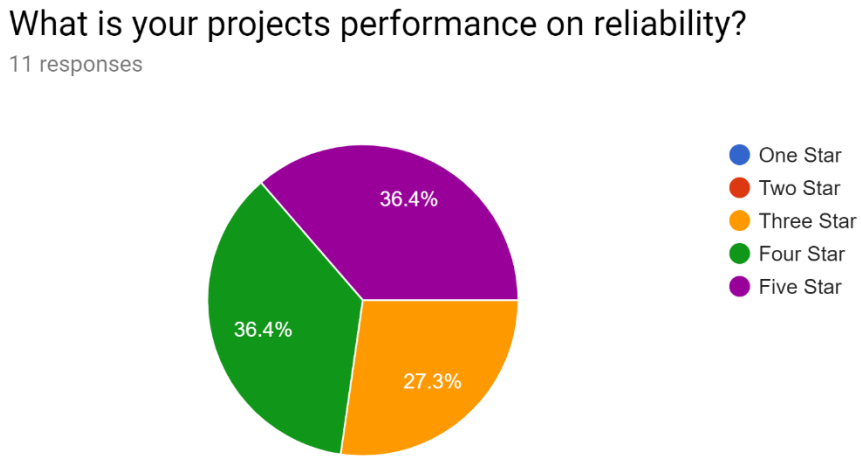


Figure 8. 3: Project Performance on Reliability pie chart

Based on the responses, almost 70% of the researchers have rated their project to reliable with an equal share of rating of 4 and 5 stars respectively. Whereas, rest 30% have shared it to be averagely reliable with a rating of 3 stars.

8.3.4 PORTABILITY

Portability refers to the usage of TDIL programme on multiple platforms. In regards to that a five-point rate scale was thus given, wherein 5 stars related to most portable and 1 being the least.

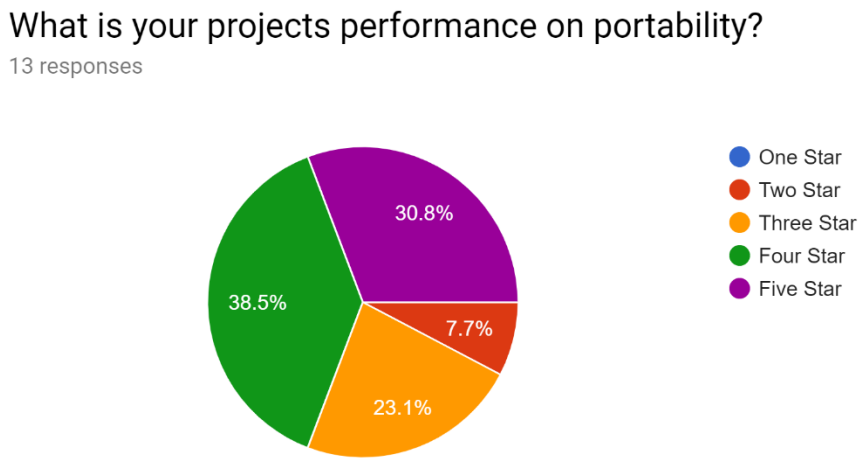


Figure 8. 4: Project Performance on Portability pie chart

Almost 70% of the respondents rated the portability as 4 or 5 stars. Whereas, almost 23% rated it 3 stars and rest almost 8% have responded the portability to be below average with a 2-star rating.

8.3.5 INNOVATION

Innovation has always been a key factor in developing and emerging technologies. Thus, keeping that in mind we proposed a question of asking the innovation and creative level of the project.

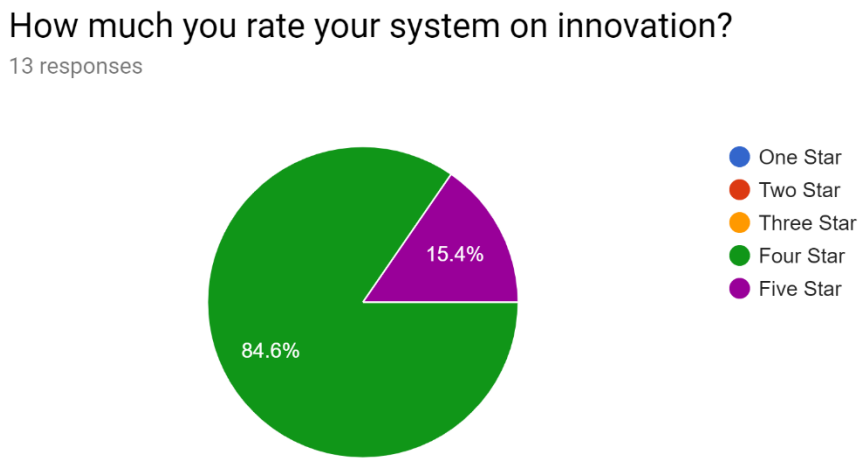


Figure 8. 5: System Innovation rating pie chart

Based on the 5-point rate scale, almost everyone has rated their system to be innovative, wherein almost 85% of the developers and research institutes have constituted it to be of a 4-star rating and rest (almost 16%) gave it a 5-star rating.

8.3.6 EASE OF INTEGRATION

Following the same 5-star rating pattern the system was then assessed upon the ease of it compatibility with other tools and services.

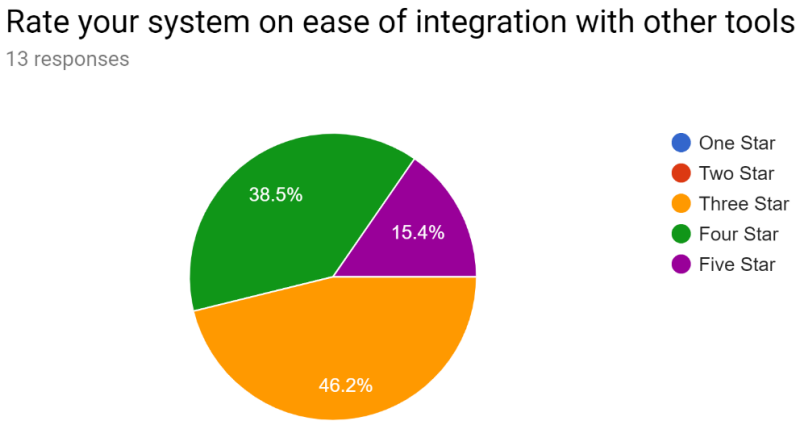


Figure 8. 6: System ease of Integration rating pie chart

The responses received suggests that the project is compatible with other tools and services, wherein, every respondent has given a rating of 3 star or more. Almost 50% rate it with 3 stars and 38% and 16% rate it 4 and 5 stars respectively.

8.3.7 PROGRAMMING SKILLS

Programming skills are the base of developing these various language technologies. To our curiosity and transparency of information we dug deeper into understanding what all programming skills were used primarily. The idea was to know the most prevalent programming language in use and to critically analyse its benefits and downfalls.

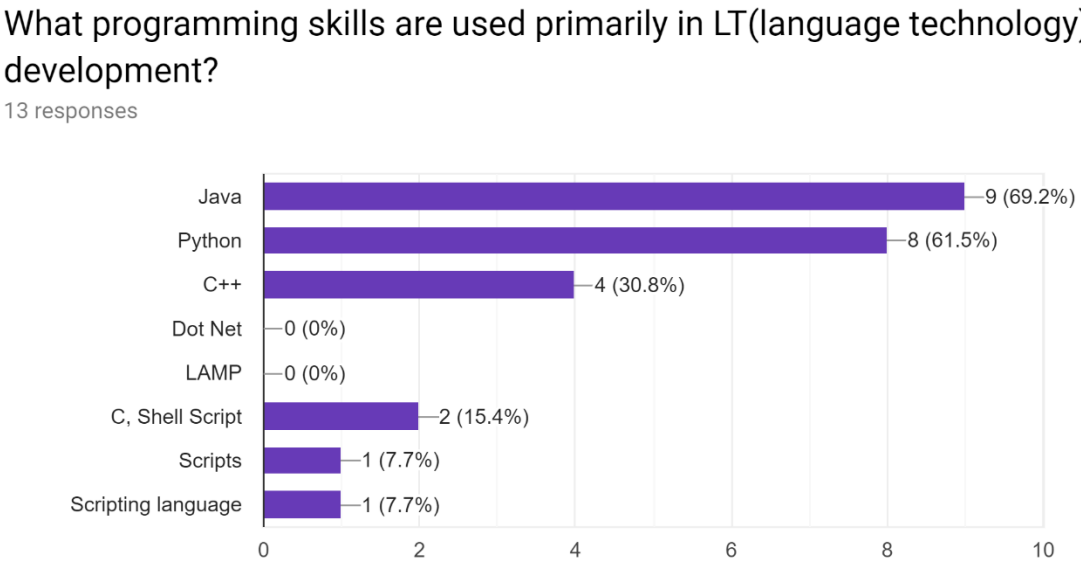


Figure 8. 7: programming skills primarily used in Language Technology

Based on our survey, we deduced that from the given options of Java, Python, C++, Dot Net, LAMP, C, Shell Script, Scripts and Scripting Language, primarily Java has been used as the language to develop this technology. Almost 70% of the developers have used Java as the prime programming language, followed by Python used by almost 62% of the developers. Other languages used includes of C++ by almost 31% and C, Shell Script by around 16% of the developers. Scripts and Scripting languages are the least used languages by the developers constituting of almost 8% each.

8.3.8 EMERGING TECHNOLOGIES

Language technology worldwide has a very innovative future if used and developed in collaboration with various future technologies. Currently, we are in the era of Industrial revolution 4.0 wherein the world is lead by AI, IoT and Blockchain. India itself has made tremendous advancements in the field of AI and in future is ready to hold the flagship of being the leader in world of technology. In order to understand this, we tried knowing about the conscience and future plans of developing Language Technology in collaboration of emerging technologies and their contribution to it.

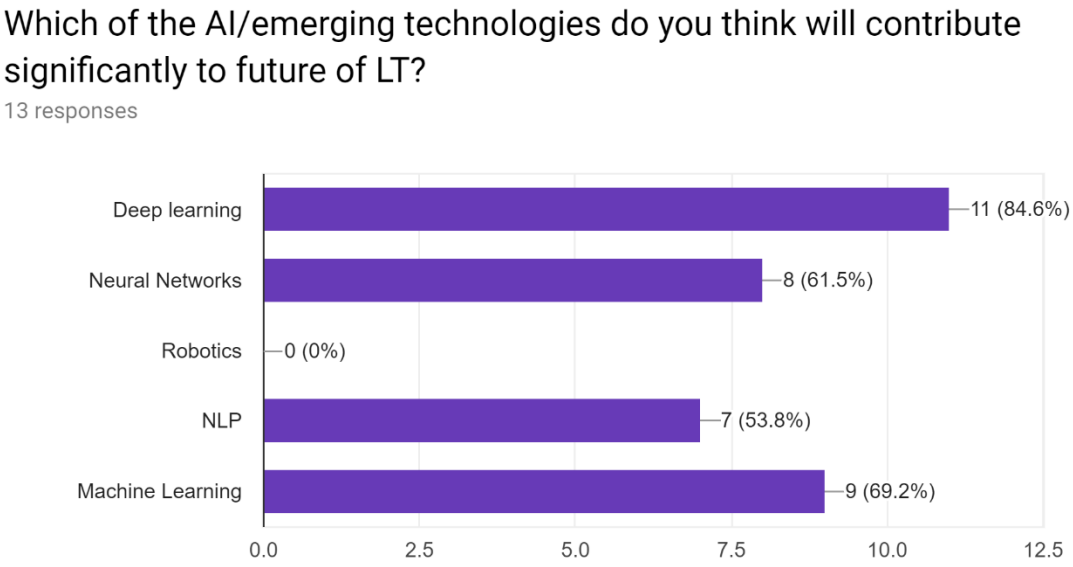


Figure 8. 8: merging Technologies with contribution possibilities in LT

Based on the above graph, it was deduced that almost 85% of the contributors believes that Deep Learning shall be a significant role player in the future of Language Technology. Differing by a slight margin, 70% observes Machine Learning as another contributing factor to language processing and tools. Furthermore, almost 62% and 54% observes Neutral Networks and Natural Language Processing, respectively, as two major contributors to the world of Language Technology. Whereas, Robotics is believed to hold no place in the future of language technology.

8.3.9 SUPPORTED DEVICES

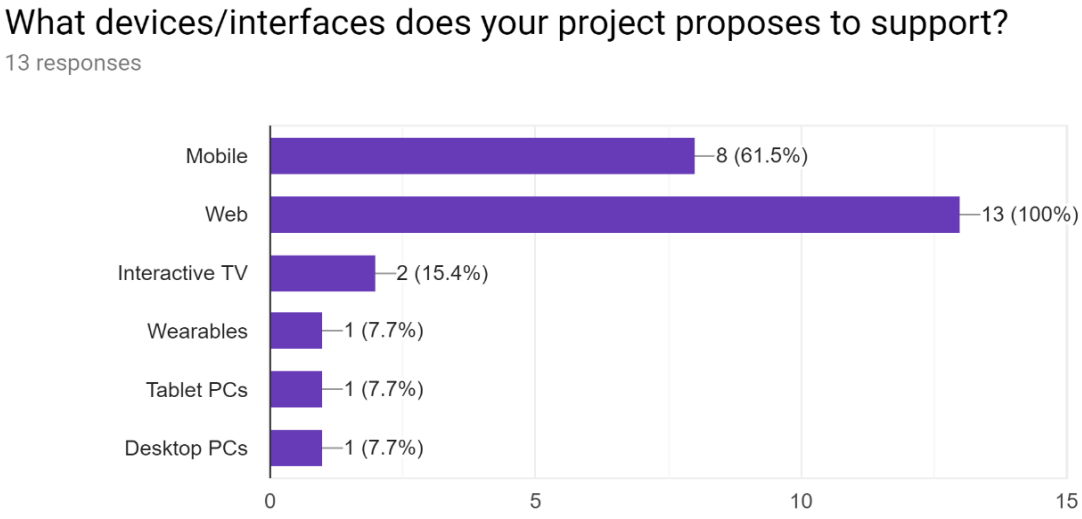


Figure 8. 9: Projected Devices/Interfaces supported by the projects

As per the data collected, web is the most compatible interface that has been focused upon, followed by almost 62% for mobile devices. Other than these, Interactive TV, Wearables, Tablets and Desktop are the least favoured ones for the development of language technology.

8.3.10 COMPARISON WITH OTHER TECHNOLOGIES

To understand the perception and status of developers in regard to the technology they have made, the question focuses on comparing the developed tools with other available tools in the market.

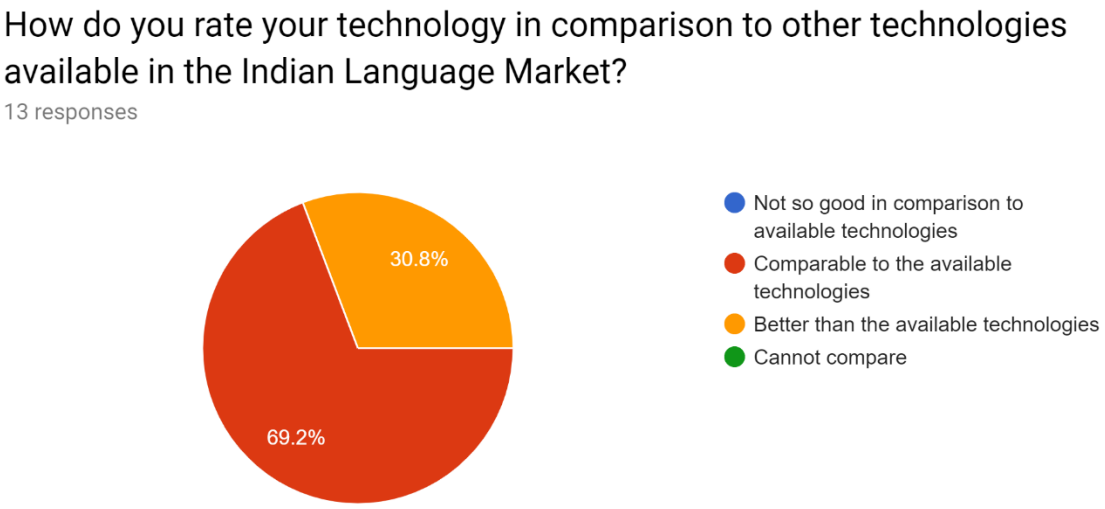


Figure 8. 10: Technology comparisons with other technologies

Based on the survey, more than half (almost 70%) of the organisations rated their technology to be efficient enough to be compared to the other available technologies. Rest 30% find their developed tools and technology to be more efficient and better than the available ones.

8.3.11 GAP IN FACILITIES FOR THE DEVELOPMENT

To understand future course of action and current setbacks, the questions were proposed based on self-awareness and knowledge of the organisations on the facilities lacking for development of Language Technology.

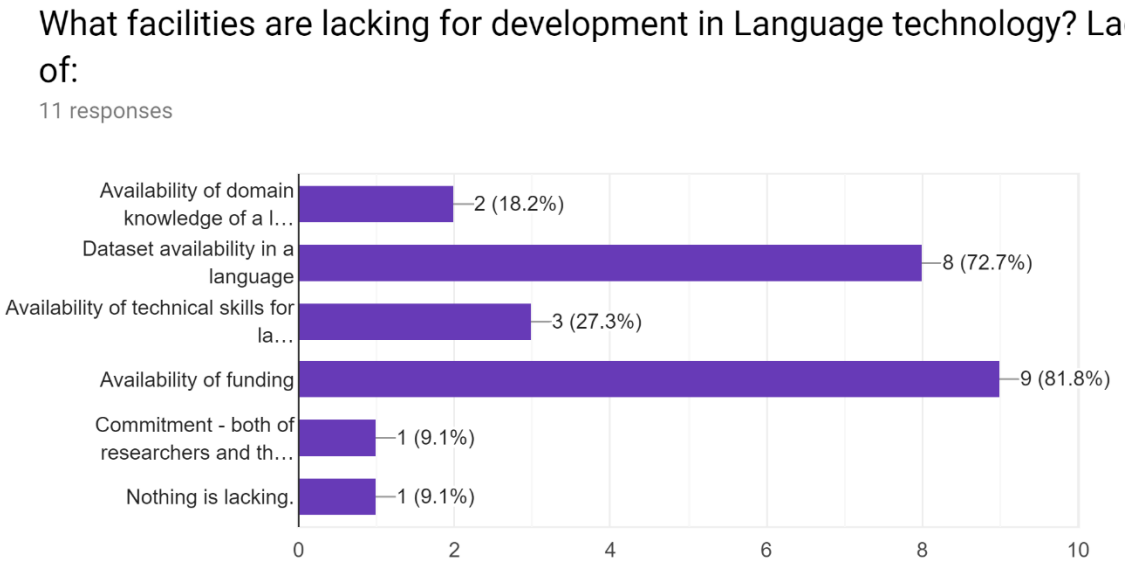


Figure 8. 11: Lack of facilities for the development of LT

Major problem or setback faced by almost 82% of the developers is in term of the funding they receive. Further, lack of availability of dataset in a language is another key problem faced by almost 73% of the developers and almost 46% face hinderance regarding the availability of domain knowledge and technical skills.

8.3.12 FUTURE THRUST AREAS IN LT

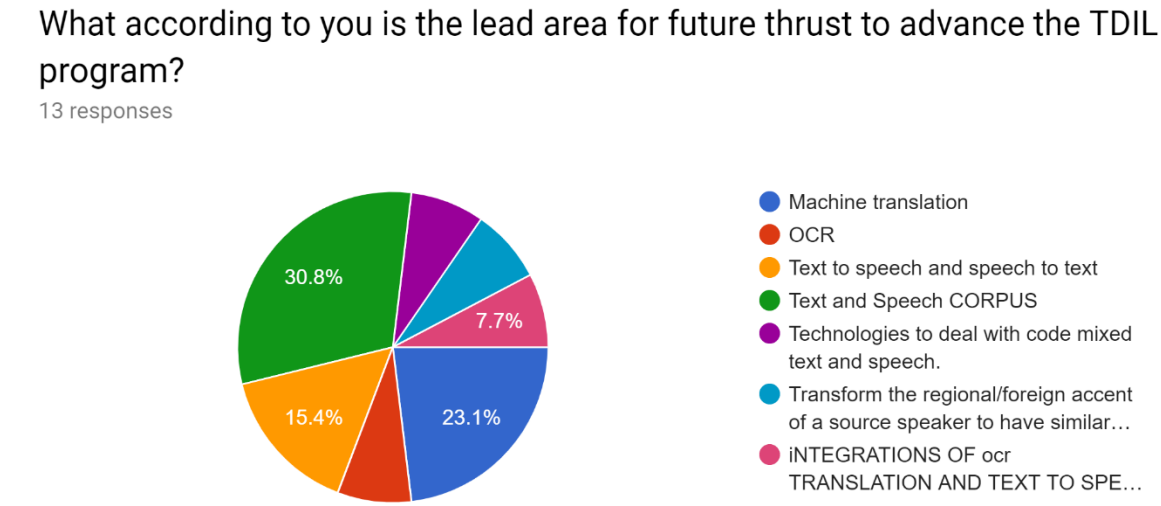


Figure 8. 12: Future thrust areas to advance TDIL

TDIL presently has a long way to go, with major impacts pending on the various sectors of the society. Regarding that a question was proposed, to know more about future interests to develop various tools of TDIL.

According to the survey, almost 30% developers find Text and Speech Corpus development as the lead area for the future thrust. Whereas, around 16% find text to speech and speech to text as the way forward for TDIL. Rest of the 30% have raised other miscellaneous areas of development for the advancement of TDIL.

Observations

Average to high on usability and efficiency. TDIL technologies are comparable or better with the existing technologies. Future thrust area is converging towards live conversations in different languages.

Developed systems are rated high on innovation. While on portability and reliability it moves from average to very high. Ease of integration looks like an area of concern. Systems readiness for the integration is lacking for a finished product point of view. Future development needs attention on more seamless integration strategies.

Availability of funding is raised as the major area of concern by most of the respondents. 100% of the projects are developed for web interfaces and around 61% supports mobiles. But support on emerging devices seems to be non-existent. Interestingly Deep Learning, Machine Learning and Neural Networks emerges as the technologies going to impact future of language technology domain. All the three are interwoven with each other. Surprisingly, only 50% respondents have put NLP as the future LT and this is the main domain under which all the development is happening in the industry.

8.4 ECONOMIC

In the course of the study, Economic parameters of TDIL has been ascertained through four important consideration which are as follows:

1. Sufficiency of Funding
2. Continuity in Release of Funds
3. Sector-Wise Benefits
4. Licensing and Suitable Business Models for LT
5. Impact on GDP

8.4.1 SUFFICIENCY OF FUNDING

Funding has been a major factor in the development of the project. Based on the economic parameters we had raised the question of availability of funds for the project.

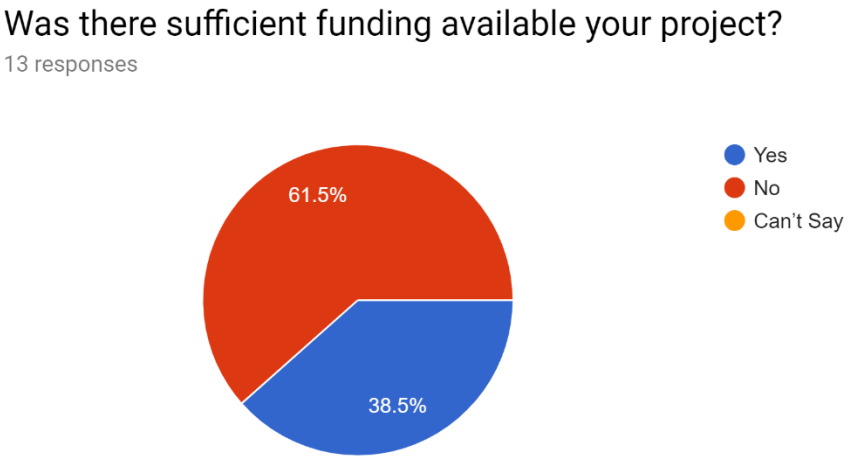


Figure 8. 13: Funding availability

Based on the survey, it was deduced that more than 60% of the respondent replied "No" to the availability of enough funds and the rest were satisfied 40% were satisfied with the funding.

8.4.2 CONTINUITY IN RELEASE OF FUNDS

Similarly, the question of release of timely funds was then put up to the organisations, developers and research institutes.

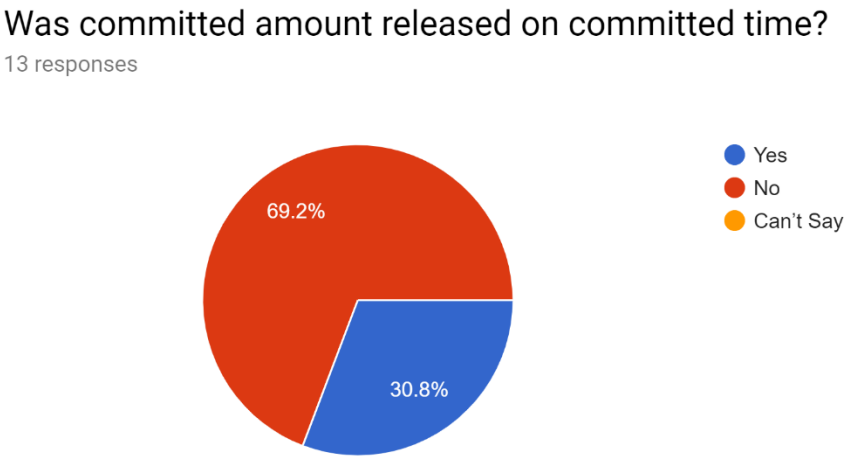


Figure 8. 14: Allocated fund release

According to the responses, almost 70% replied with a "No" to timely and continuous funding the project whereas rest 30% were satisfied with the timing and commitment of the funds.

8.4.3 SECTOR-WISE BENEFITS

In order to understand the impact of TDIL on various sectoral growth the question of sector-wise benefits was then put up.

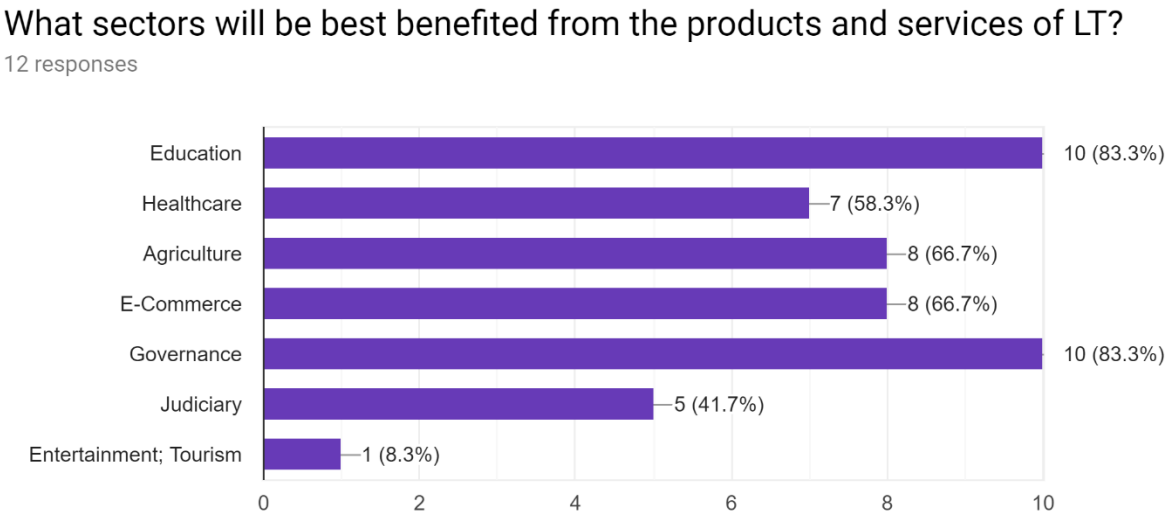


Figure 8. 15: Sectors projected to benefit from LT

As per the responses, almost 90% had voted for education and governance sector to have the most impact. Whereas, 80% respondent also agreed upon the impact on e-commerce and agriculture and rest 58% and 42% also found healthcare and judiciary, respectively, to also be a part of the benefits of TDIL.

8.4.4 LICENSING AND SUITABLE BUSINESS MODELS FOR LT

In order to gain a better insight into the conscience of business management that is supposedly being followed. And how the commercialisation of TDIL takes place, the question was thus asked about the technology licensing in the development of language technology.

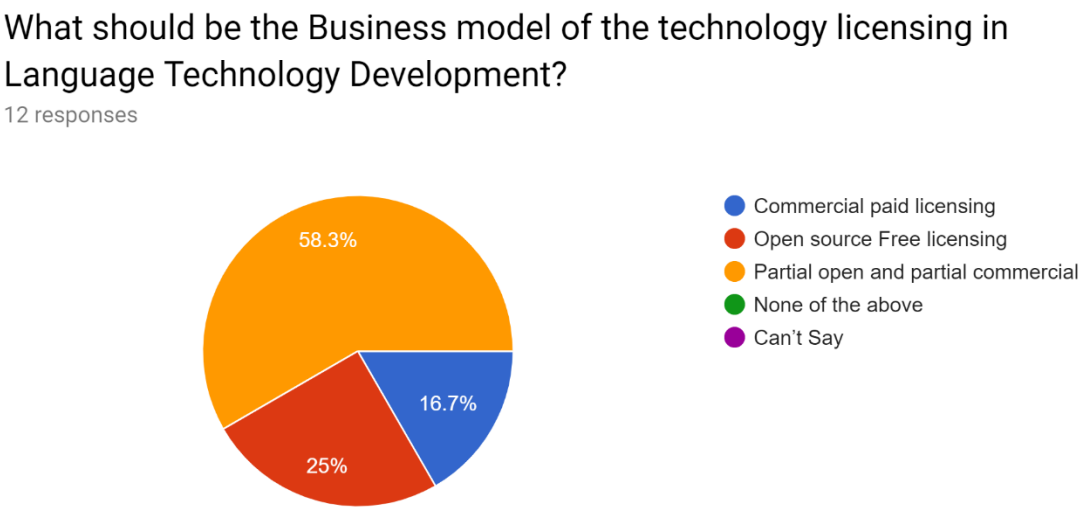


Figure 8. 16: Business model for technology licensing

Major responses, accounting for almost 60%, were in favour of keeping the licensing partially open and partially commercial. Whereas rest 42% were divided into two sets of keeping the licensing commercialised paid or open source free, wherein 25% voted for open source and rest 17% voted for paid.

8.4.5 IMPACT ON GDP

Since TDIL is supposed to and believed to be quite impactful for various sectoral growth, it therefore, shall have a healthy impact on the GDP. Based on this, the question of impact on GDP was proposed.

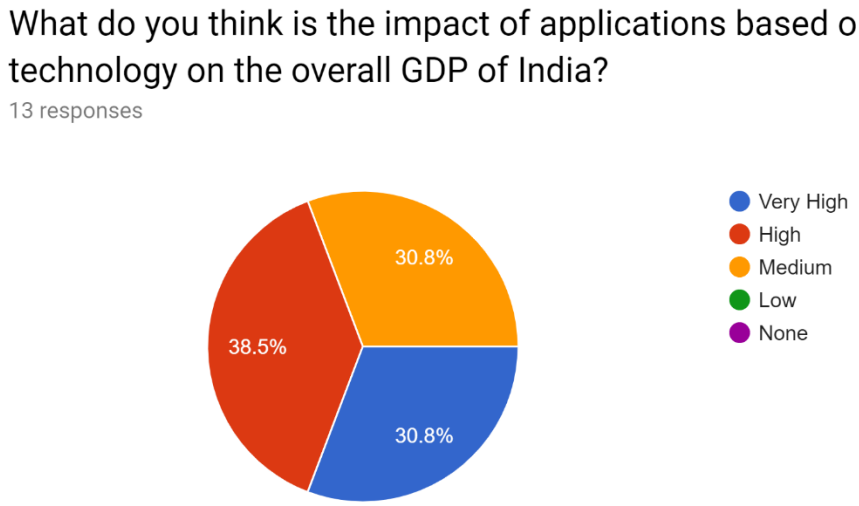


Figure 8. 17: T impact on GDP of India

On a variation of none to very high, the respondents were asked to choose on the of following options from no impact or none, low impact, medium, high and very high impact of TDIL on GDP. Based on the responses, 70% feels TDIL will have a huge impact on the GDP, whereas, the rest 30% feels the impact to be medium and not as effective as to what the rest of the respondents feel.

Observations

Funding was neither sufficient not consistent which was raised by the technical teams in the earlier section also.

Education and Governance are the biggest sectors that can be benefitted from the LT as per the technical teams. Next in order are Agriculture, e-Commerce and Healthcare sectors which will reap the benefits of TDIL.

It will have an immense impact on the GDP of the country according to most of the technical teams.

8.5 OPERATIONAL

1. Lack in Skills
2. Readiness for Deployment
3. Readiness for Commercialisation
4. Satisfaction with Deployment and Hosting Infrastructure

8.5.1 LACK IN SKILLS

Based on the operational and organisational skills of the project, the question was then proposed of the lack of skills as per the research organisations and institutes.

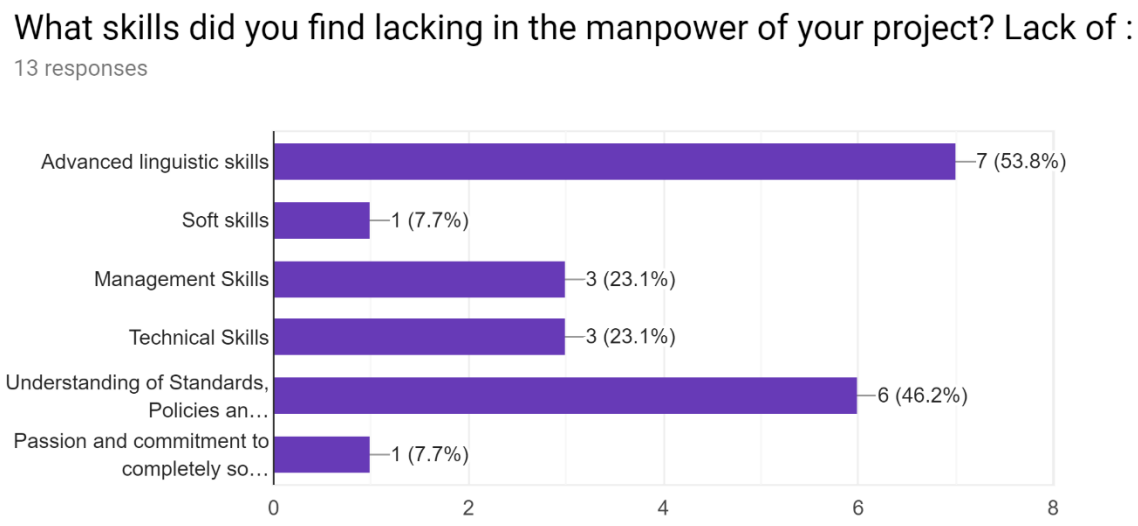


Figure 8. 18: Skills lacking in manpower of projects

As per the survey done, it was found that majority of the institutes and organisations found advanced linguistic skills lacking in their manpower, constituting of almost 54%. Whereas, other 45% found Understanding of standards and policies to be lacking and rest almost 46% found their manpower to be lacking in management skills and technical skills, respectively. Other than that, soft skills and passion towards the project were also the skill set to be found missing.

8.5.2 READINESS FOR DEPLOYMENT

In order to understand the readiness of the project for the market, based on a percentage rating scale the research institutes and organisations were asked to scale out the readiness for deployment of the project.

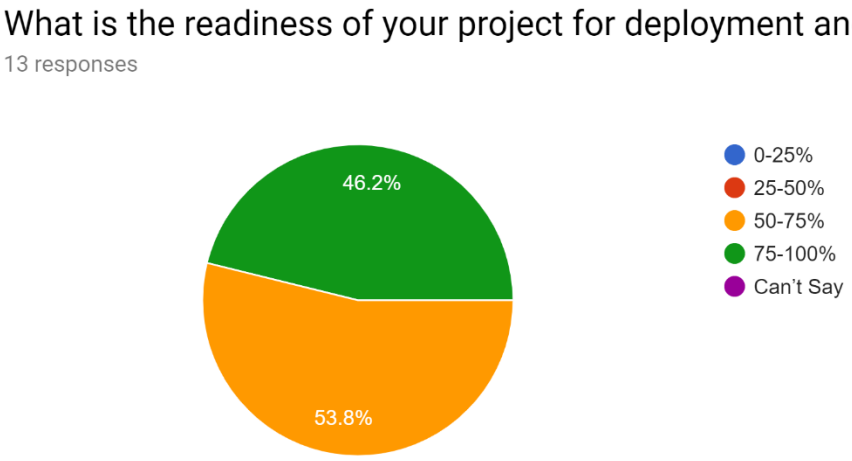


Figure 8. 19: Project redness for deployment and usage

Everyone rated their project to be ready for deployment, wherein almost 47% range between the scale of 75-100% and rest 54% (rounded off) range between 50-75%. Thus, showing us the already ready and prepared mindset and readiness of the institutes and organisations for the deployment of the project.

8.5.3 READINESS FOR COMMERCIALISATION

To collate more information about the readiness of the project in terms of market accessibility and commercialisation process of the project. On a percentage scale of 0-100% and Can't Say, the question of readiness of commercialisation was then proposed.

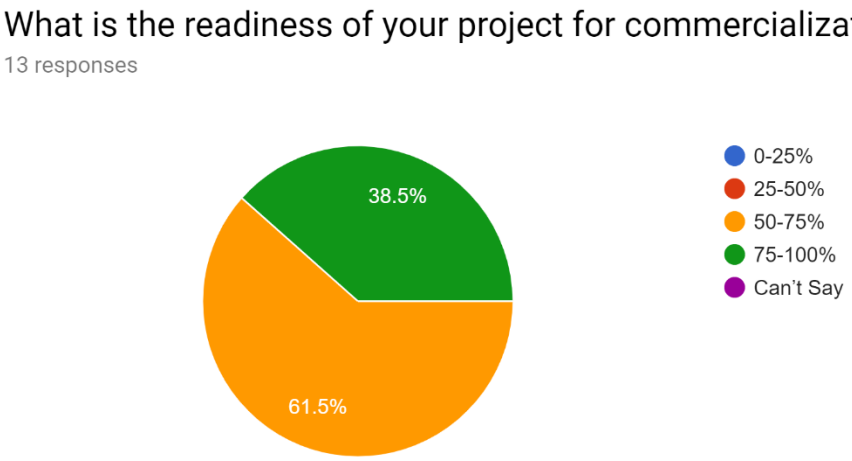


Figure 8. 20: Readiness of project for commercialisation/market

Based on the responses, averagely everyone believes in the readiness of their project for commercialisation process. More than 60% find it to be ready for commercialisation thus putting themselves in the range of 50-75% and rest 40% are confident enough with their preparedness, thus, putting themselves in the range of 75-100%.

8.5.4 SATISFACTION WITH DEPLOYMENT AND HOSTING INFRASTRUCTURE

Based on the personal satisfaction of the deployment of the TDIL project and its hosting infrastructure the research institutes and organisations were asked to rate the services on a 5-point rate scale. Wherein, 5 stars being the most satisfied and 1 star being the least satisfied.

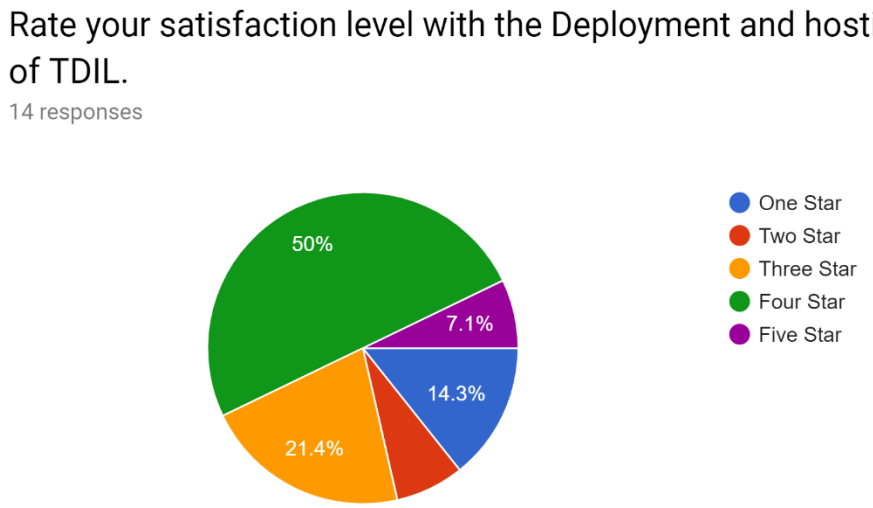


Figure 8. 21: Satisfaction with deployment and hosting infrastructure

Based on the analysis and the responses received the ratings varied from being a 5-star satisfaction to as low as 1-star satisfaction. Nearly, 8% find the services to be most satisfactory with a rating of 5-stars. Majorly, 50% find the deployment and hosting infrastructure of TDIL to be satisfactory enough to give it a 4-star rating. Whereas, rest nearly 22% and 14% find the services to be averagely and below satisfactory thus giving them a rating of 3-stars and 1-star respectively.

Observations

Advanced linguistic skills and policy & standards emerges as the skills team feels are in sparse.

As per the research teams, the readiness for deployment and commercialisation is 50% to 100%. The systems have been deployed on DC platforms as well.

As far as deployment is concerned, they lack the finishing touch to be used as a commercial product. Efforts on the commercialisation front seems to be very low as well. Deployment also needs a high level of scale up because current uses are too thin to make any impact on the technology use by masses.

It is also imperative that there be annual review of the programmes based on their user impact and need analysis. This would enable timely intervention where a push is needed (refer to Chapter 10.4.5)

8.6 SOCIAL

1 CAPACITY BUILDING IN SOFTSKILLS

2 AWARENESS ABOUT LT

3 IMPACTING BUSINESS AND JOB OPPORTUNITIES

8.6.1 CAPACITY BUILDING IN SOFTSKILLS

This question was asked to understand the softskill training requirements of the area that can affect the language technology development.

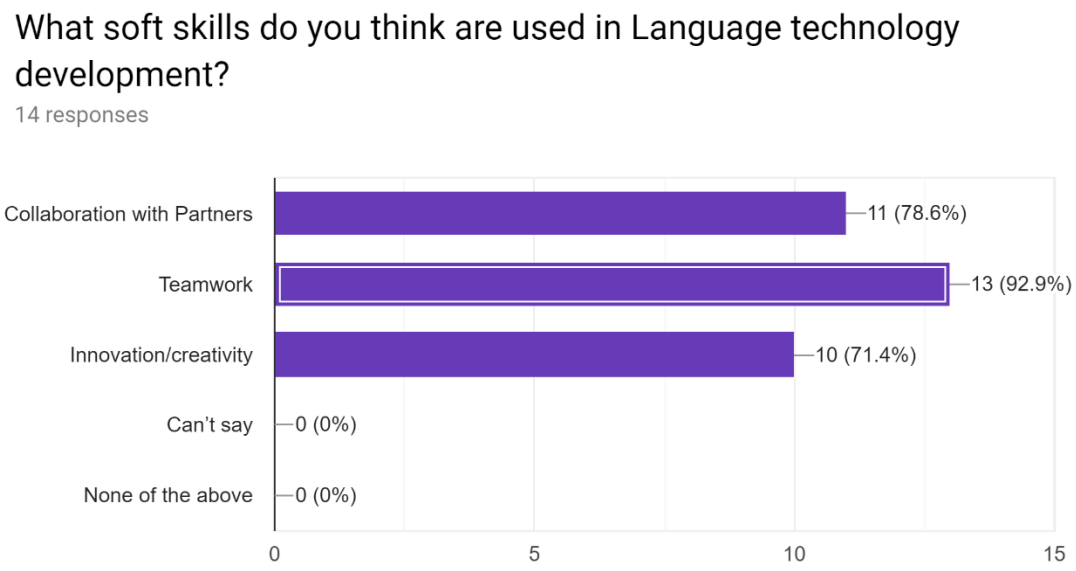


Figure 8. 22: Soft skills used in LT

Based on the responses, nearly 93% voted for teamwork, followed by 79% voting for collaboration with partners and other 72% voted innovation and creativity as the required soft skills.

8.6.2 AWARENESS ABOUT LT

For any project to have a social impact, the basic requirement is of the society to be aware about the initiative. Keeping this mind, the question was proposed to ask the level of awareness as per the research institutes and organisations about language technologies by the society.

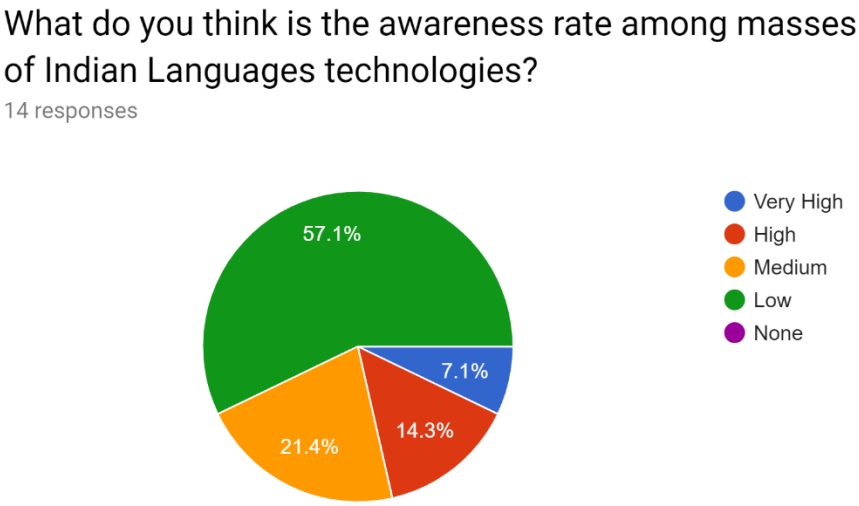


Figure 8. 23: Awareness amongst masses on availability of Indian LT

Ranging from a scale of Very high to None, the respondents majorly accounting for almost 58% replied as the awareness level to be low. Whereas 38% rated as the awareness level to be fairly satisfactory and other 7% find the awareness to be relatively higher.

8.6.3 IMPACTING BUSINESS AND JOB OPPORTUNITIES

To understand social impact in a better light, IIPA researchers proposed the question of impact of Language Technology on the creation of various businesses and job opportunities. On a scale of very high to can't say, the respondents were asked to choose one and reply.

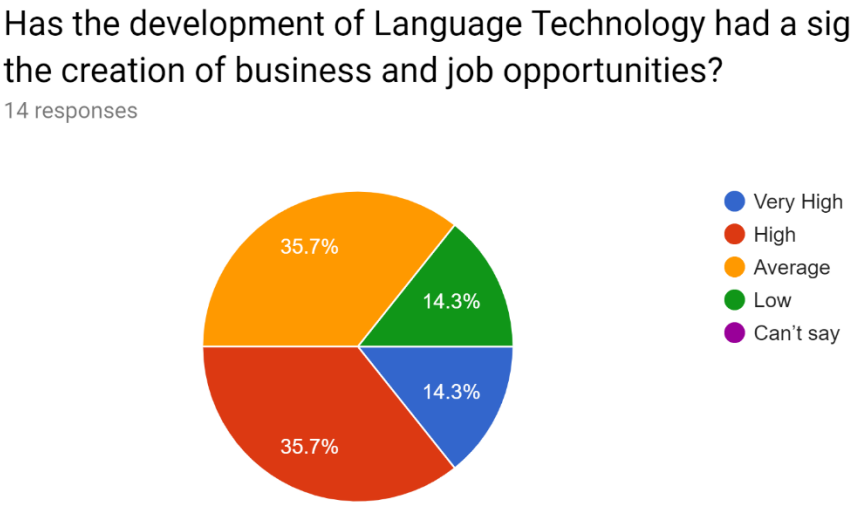


Figure 8. 24: Impact of LT on business and job opportunities

Based on the survey it was observed that the responses could be easily dived into equal halves. Wherein, one share of 50% respondents believes that the impact of Language Technology is relatively higher and thus quite impactful, whereas, rest 50% find the impact to below average or low, respectively.

Observations

According to the technical teams, the awareness about language technology among masses is appallingly low.

LT will have a high impact on the increase in job and business opportunities for people.

As far as soft skills are concerned, they thin teamwork is the most important soft skill, followed by collaboration, innovation and creativity.

8.7 SUMMARY

A survey was conducted that included researchers from different TDIL projects. It made a few findings:

- 94% of respondents found their projects user-friendly.
- All respondents found their projects at least passably efficient.
- 70% found them very reliable.
- 93% found them at least moderately portable.
- All respondents found their projects innovative (85%) or highly innovative (15%).
- All respondents found their projects easily compatible with other tools and services.
- Java was the most valued skill in TDIL projects, followed by Python and C++.
- All developed applications had web versions. The second most supported device for most applications was the smartphone.
- All respondents felt their technologies were comparable to (69.2%), or better than (30.2%), their alternatives in the IT market.
- 81% of respondents said that TDIL faced a shortage of funds.
- 72.7% of respondents said that TDIL did not have as large a linguistic dataset as was ideal.
- 27.3% of respondents said that the required technical skills were in short supply.
- 30.8% felt that the expansion of the text corpus was the best future step for TDIL, while 23.1% felt that machine translation was the way forward.
- 61% of respondents felt their projects had not received adequate funding, and 69.2% felt their funds had not been provided regularly or on time.
- The sectors most likely to benefit from TDIL, according to the respondents, are education, governance, agriculture and e-commerce.
- 58.3% of respondents agreed that the ideal business model for the language technology industry was one in which technology was partially open source and partially commercial. Of the remaining respondents, 25% felt that language technology should be completely open source.

- All respondents thought that language technology-based applications had a considerable impact on GDP.
- 53.8% of respondents felt there was a lack of advanced technical skills in their project teams. 46% felt their teams had a less-than-perfect understanding of standards and policies.
- The majority of respondents (53.8%) said their projects were at 50-75% ready for deployment, while the remainder said that their projects were 75-100% ready.
- Almost two-thirds of the respondents said their projects were moderately ready for commercialisation. The remainder said that they were almost completely ready.
- About 78% of the respondents were at least fairly satisfied with the deployment and hosting structure of TDIL.
- According to the survey, the number one soft skill used in language technology development was teamwork, followed by collaboration with partners and innovation.
- More than half of the respondents felt that the awareness of TDIL initiatives among the masses was low.
- Half of the respondents felt that language technology had a high impact on the creation of new business opportunities, while the other half felt it had an average or low impact.

Chapter 9: Comparison with Similar Initiatives

9.1 INTRODUCTION

India, along with several other nations, has taken steps towards safeguarding its languages from digital obsolescence. To this end, the governments of these countries and several private corporations have made efforts to build multilingual tools for digitisation and digital preservation of heritage and culture. From the Canadian government to Microsoft Corporation, entities everywhere have acknowledged that unless the existing information and communication technology is made available in regional languages, the internet – and the rest of the digital world – will remain out of reach for the masses.

Besides fulfilling the aim of preserving the precious language resources, developing language technologies also accomplish eclipsing the digital divide that is increasingly plaguing development in several countries. This helps individuals who often lose out in the competition for development because their first language is often not the digitally dominant language. Hence, initiatives, around the world, developed for bringing endangered, indigenous, and minority languages to mainstream, not only help in making these languages available for future generations, but also safeguard them from obsolescence and promote global cooperation and free sharing of information.

TDIL or Technology Initiative for Indian Languages is one such initiative by the Government of India, which has taken the step towards building language technologies for Indian languages. Similar to TDIL, around the world, several governments and private organisations, namely, Africa, Canada, European Union, Microsoft, Google, etc., have also launched such programmes. These initiatives have developed various tools and incorporate several stakeholders to fulfil the aim of safeguarding languages while promoting their development. The following chapter looks at comparing these language technology initiatives to understand the extent to which the technologies that have been developed and utilised. This study will also try to benchmark TDIL amongst other initiatives to analyse the plausible areas where TDIL is performing well and areas where it is lagging. The chapter will end with the author recommending TDIL of plausible areas of improvement.

9.2 STUDY METHODOLOGY

The benchmarking process of TDIL has been undertaken by following the below mentioned steps:

Selection of global best practices

For comparison of TDIL with the global best practices, the first step was to select these practices. Extensive review of literature along with several brainstorming sessions were done to decide these. The team settled on comparing four initiatives, namely, African Language Technology Initiative (Alt-I), Canadian Indigenous Languages Technology Project, European Union Language initiatives, Google Translate and Microsoft Translate.

Understanding of the initiatives

The websites of each of these initiatives were then critically studied to understand each of the initiative's history, individual initiatives, etc. The discussions with the team also helped in gaining an insight into each of the initiatives.

Comparison of initiatives

Post having a macro level understanding of each of the initiatives, an unbiased classification of each of the platform was undertaken on select parameters based on secondary research, namely:

- Name of the entity
- Type of entity – public, private or other
- Year of founding
- No. of target languages worked with by the entity
- Services provided to beneficiaries
- No. of beneficiaries
- Agencies involved/constituents of managing committee

- Funding
- Key initiatives.

9.3 DIFFERENT GLOBAL INITIATIVES

9.3.1 AFRICAN LANGUAGE TECHNOLOGY INITIATIVE (ALT-I)

Introduction

Setup in 2002, the African Language Technology Initiative (Alt-I) is a non-profit organisation that works on making African languages compatible with contemporary information and communication technologies. In order to popularise the best practices for the research and development of African languages, it develops software and hardware, undertakes relevant studies and regularly disseminates its findings[14].

Structure and Funding

Alt-I is funded by international grant-making organisations such as Bait Al-Hikmah, OSIWA, IDRC, and ACALAN. Alt-I's governance is built on a two-tier structure consisting of an Advisory Board and an Executive Management Team. The general strategic guidelines for the activities of the organisation are set by the Advisory Board, which includes eminent scholars with expertise in language, linguistics and engineering[15].

[14] African Language Technology Initiative.(2002). In *ALT-i*website. Retrieved September 11, 2019 from http://www.alt-i.org/?page_id=9

[15] Grover, S.A., Huyssteen, B. G., Pretorius, W.M. (2011). The South African Human Language Technology Audit. *Lang Resources and Evaluation*(45), 271-288. Retrieved from https://link.springer.com/article/10.1007%2Fs10579-011-9151-2

Keys Areas of Research

Speech Recognition

Many of Africa's languages are tonal, meaning that the meanings of words rely as much on their musicality as on their phones. Alt-I's speech recognition initiative is a novel attempt to recognise African tone languages using Yoruba as a pilot language. Its approach is based on the use of tone information to increase the speed and accuracy of the recognition process.

Machine Translation

Alt-I is working on machine translation systems to allow the translation of texts between two language pairs: Igbo-English and Yoruba-English.

Corpus Development

The team is developing a functional corpus of computer readable Yoruba texts in standard orthography and a statistical language model of Yoruba.

Activities

Alt-I has pioneered several key projects and initiatives for the purpose of making ICT technology accessible to the speakers of African languages. Some of these have been listed below:

- *L10N*

The L10N project focused on the localisation of Microsoft Windows and Microsoft Office for African users, adding support for the three most spoken languages in Africa: Hausa, Igbo and Yoruba. Alt-I conceptualised and executed this project in collaboration with the Microsoft Corporation.

- *Redefining Literacy*

Funded by OSIWA (Open Society Initiative for West Africa), Alt-I's Redefining Literature programme aims at using speech recognition and speech synthesis technology to make literature available to people who cannot read nor write.

- *Advocacy*

Through the Nigerian Community Radio Coalition, Alt-I is actively promoting the use of community radio broadcasting as a means of vitalising minority languages and documenting oral

texts. This is an important initiative in Africa, where there is a considerable danger of losing cultural capital to Westernising forces.

- *Engagement with universities*

The initiative is working with Nigerian universities to develop a crop of multidisciplinary programmes which in turn produce researchers who are experts in the field of linguistics, engineering and computational sciences.

9.3.2 CANADIAN INDIGENOUS LANGUAGES PROJECT

Introduction

The Canadian Indigenous Languages Technology Project is an undertaking managed by the Government of Canada. This project aims to assist indigenous language educators and students by developing speech- and text-based technologies for the stabilisation, revitalisation and reclamation of Indigenous languages. It does so by enlisting the services of Indigenous language translators, transcribers and other language professionals. Since the majority of texts in Native American languages are oral, it greatly promotes the accessibility of audio recordings. The technology it develops is released to indigenous communities as open-source software.

Structure and Funding

In its 2017 budget, the Canadian government had allocated $89.8M for the development of Indigenous cultures over three years. The Canadian Indigenous Languages Technology Project, managed by the National Research Council (NRC), received $6M of this amount[16].

Key Areas of Research

Text Technologies

Under the project, keyboard layouts for many indigenous languages have been made available. These have incorporated predictive methods to suggest complete words and phrases that are about to be typed, encouraging young users as well as second-language learners. The project

[16] Government of Canada. *Canadian Indigenous Language Technology Project 2007.* Retrieved September 8, 2019 from https://nrc.canada.ca/en/research-development/research-collaboration/programmes/canadian-indigenous-languages-technology-project

has also developed machine translation, spell checking, and orthography conversion technology to facilitate the conversion of text in one orthographic form to another.

Speech Technologies

The project has facilitated speech recognition that can be used in keyword-based searches. It works around agglutination using inbuilt audio-segmentation and speaker diarization in its software's. It has also worked on text-to-speech technology.

Image Technologies

The project has developed optical character recognition for Canadian aboriginal syllables and Roman orthographies.

Computer-aided language learning (CALL)

Computer-aided language learning (CALL) course modules are widely available for indigenous languages, particularly through the FirstVoices Language Tutor (FVLT) portal, which offers approximately 50 online courses covering many indigenous languages. These consist of exercises on listening, speaking, reading, and vocabulary development, as well as online language-learning games. There are even basic phonetic tutorials focusing on phone acquisition and familiarisation.

Activities

- *The East James Bay Recordings*

The Canadian Broadcasting Corporation (CBC) creates programming by and for indigenous people, providing services in eight Native and Inuit languages. As part of the NRC's indigenous languages technology project, the CBC has provided the Computer Research Institute of Montreal[17] (CRIM) with access to East James Bay Cree recordings. Its audio processing technology efficiently indexes the spoken content of very large audio databases, making this content easily accessible through search engines. CRIM's speaker recognition technology, which can identify the identities of the people who generated specific speech segments, is world-class. With access to the East James Bay Cree recordings, CRIM has the potential to develop audio segmentation and analysis tools suitable for indexing audio recordings in indigenous languages which is a major step towards documenting Native American culture and history.

- *Pirurvik*

[17] The Computer Research Institute of Montréal (CRIM) is an applied research and expertise centre in information technology. Its speech and text team have a long and distinguished record of accomplishments in speech recognition technology.

Pirurvik is a centre of excellence for Inuit language, culture and well-being. The main focus of the NRC's collaboration with Pirurvik is the transcription of spoken Inuktut drawing on audio recordings. Pirurvik selects audio material that is in pure Inuktut and contains a depth of vocabulary.

- *Creating Online Indigenous Language Courses (COILC) initiative*

The NRC is a partner of the experts of 7000 languages, a US-based non-profit, non-indigenous organisation that creates courses for endangered languages around the world. It has declared that it will fund selected community teams who wish to create online courses for their languages.

9.3.3 EUROPEAN UNION (EU): LANGUAGE TECHNOLOGY FOR MULTILINGUAL EUROPE

Introduction

Multilingualism is core to the idea of Europe and language technology initiatives are a high priority in the region as it is the centre for various multilingual communities. It has 24 official member state languages, and is home to many unofficial and regional languages, as well as languages of minorities, immigrants and important trade partners. While Europe's multilingualism is often celebrated, it is also acknowledged that language barriers severely hamper the free flow of information, ideas, goods and services through the continent. In an attempt to foster collaboration and create more cultural awareness for a strong and united Europe, artificial intelligence, deep learning and high data sets, when used with language technologies like neural machine translation and statistical translation, can bridge the gap and create digital harmony. This will further improve the socio-economic integration of the continent and lead to a stronger Europe.

Structure and Funding

As far as funding is concerned, a shared responsibility between the European Union, industry and member states was envisioned with the EU as the stakeholder that should be "naturally" responsible. The distribution of votes for stakeholder involvement looks as follows: European Commission (89%), Industry (57%) and Member states (57%)[18].

[18]Cracker Project.(2015). *Language Technologies for Multilingual Europe* (Version 1.0). Retrieved September 10, 2019 from http://cracker-project.eu/sria/

Key Areas of Research

Cross Lingual Big Data Language Analytics

It aims to develop language technology that can compute huge amount of data from various languages. It will encourage both experts as well as non-expert stakeholders from across the EU to indulge in a participatory democracy and stay informed about various changes and developments.

High Quality Machine Translation

This aims to counter the barriers that prohibit high-quality translations. It relies on translation professionals and novel quality metrics and uses human annotations for improving models.

Knowledge Technologies

The aim is to bring together all the information and data that is used by and created for understanding, translation, curation and generation. It will include knowledge graphs, linked data sets and ontologies as well as services for building, using and maintaining them.

Conversational Technologies

The aim is to create voice-controlled interfaces with multilingual conversational capabilities.

Activities

- *Human Language Project (HLP)*

It is a programme that aims to develop monolingual, cross-lingual, multilingual, multimodal, context-aware, culture-aware and knowledge-rich computational models that process human language in a precise and semantically nuanced manner. The primary goal of HLP is to solve the problems of cross-lingual communication by tackling the problems of Deep Language Understanding by 2030. It attempts to maximise the impact of language technologies on the European economy and society by initiating a long-term research, development and innovation programme, involving both innovation as well as commercialisation.

- *CRACKER Project*

The CRACKER initiative is a federation of projects and organisations working on technologies for a multilingual Europe. CRACKER pushes for an improvement of machine translation research in the areas of efficiency and effectiveness by learning from other disciplines where collaborative efforts - guided by interoperability, standardisation, common challenges and comprehensive success metrics - have led to otherwise unimaginable breakthroughs. The nucleus of this new research, development and innovation strategy towards high quality machine translation is the group of projects that will be supported by CRACKER.

- *Digital Single Market (DSM)*

It will include the sharing of data, technology, culture, language, literature and products within the European Union. The main areas of interest in the formation of a DSM are multilingual e-commerce[19], multilingual e-learning[20], multilingual e-health[21] and multilingual e-governance.

- *Digital Language Diversity Project*

The aim of this project is to bolster the sustainability of Europe's regional and minority languages in the digital world by empowering their speakers with the ability to create and share digital content in their native languages. This project promises significant benefits for European society and business through firstly increasing Europe's digital language diversity, secondly it will form a global community of digital content producers, activists, technicians and policy makers, who will revitalise regional and minority languages and hence assert the cultural autonomy of different nations and thirdly provide software developers, SMEs and industries with state-of-the-art products and services that allow the use of regional and minority languages on digital devices.

- *EuroMatrix*

This was an ambitious project of EU which was implemented in three years (2006-09). It sought to develop software capable of automatically translating between the European Union's 24 official languages. It was developed in order to help governments, businesses and citizens communicate more easily and economically. The team of EuroMatrix team had developed software that allowed the host computer to learn from past work and experience and hence translate more accurately. In essence, this approach relied on the program referring to an existing body of translated text and using statistical analysis to determine how different words were used.

- *SUMMA Project*

Under the Cracking the language barrier initiative, the SUMMA project developed an editorial and high-quality editing tools for analysts and journalists using language technologies such as speech-to-text, machine translation, natural language processing, text to speech, etc.

- *FREME Project*

Started in February 2015, this project aimed at providing both semantic as well as multilingual enrichment to digital content through authoring and publishing multi-lingually and semantically enriched e-Books, integrating semantic enrichment of multilingual content in the processes of

[19]Which can translate online retail sites and their products into desired languages and help SMEs penetrate the Digital Single Market

[20]Used to design learning programmes for people who wish to learn new languages which can help with self-studying, cross-border migration, training for staff members of pan-European companies, etc.

[21]This process will include improving data accuracy, not only in terms of sharpening the translatability of medical concepts, but also in terms of standardising the terminology that medical discourse produces in various EU languages.

translation and localisation and enhancing cross-language information sharing (e.g., by providing open access to agricultural and food data)

- *ModernMT*

This is a multilingual translation software provided as SaaS (available online via subscription) for business enterprises. It currently supports 45 language pairs. It is an Open Source software which can be downloaded from the GitHub repository. ModernMT works with the help of baseline models built on billions of words of premium data. It provides software customisation, RESTful API and dedicated customer service.

- MutliJEDI

The project had two main objectives: the creation of large-scale lexical resources for dozens of languages, and the enabling of multilingual text understanding. The first stage of the project aimed at developing a methodology for automatically creating a large scale, multilingual knowledge-base[22] and the second stage used this lexical resource to jointly perform disambiguation across multiple languages and to design and experiment novel graph-based algorithms.[23]

9.3.4 GOOGLE TRANSLATE

Introduction

Google Translate is a free multilingual machine-based translation software developed by Google that supports over 100 languages. It offers translation services for various kinds of input: text to text, words on an image to text, direct translation of a document, speech-to-speech translation, handwriting translation, website translation and mobile app translation. Earlier, Google Translate had relied on statistical machine translation, a method that uses rule-based predictive algorithms to find suitable translations - first for phrases, and then for sentences. In 2016, Google transitioned to use neural machine translation, which uses deep learning techniques to translate whole sentences at a time and generally provides greater accuracy.

[22](BabelNet, http://babelnet.org),
[23](Babelfy, http://babelfy.org).

Figure 9. 1: Developing an "interlingua" (credit: Google)

Structure and Funding

Google is a private Company that develops language technologies to disseminate its internet-related services.

Research Areas

Google Translate has been conducting research in the following areas:

Image to text translation

It enables the user to convert the text in the world around him into any of the languages supported by Google translation.

Text to text translation

It uses Google Neural Machine Translation that translates whole sentences rather than phrases or words and supports more than 105 languages.

Speech to Speech translation

It enables one to record the voice and translate it into a foreign language.

Document translation

The text in a document can be converted into its translated result in a foreign language.

Website and mobile-application data translation

The google converts whole webpages and results into another language for the user.

Key Initiatives

AutoML Translation

This feature is available as a paid service. It allows developers, translators and localisation experts who have limited machine learning expertise to create translation models that are high-quality and production ready. After uploading the details of each pair of languages to be translated, AutoML Translation creates a custom model which can be scaled and adapted to meet domain-specific requirements.

Translation API

This is a paid feature that can be used by website and app owners to convert their content into more than a hundred languages using Google's neural translation methods. It allows one to create custom translations for all language pairs recognised by Google.

Google Translator Toolkit

This is a web-based application which allows human translators to edit the translations automatically generated by Google Translate. Translators can upload, share and translate documents or Wikipedia articles onto the toolkit service. They can also access shared translations, glossaries and translation memories.

Google Translate Community

Google constantly seeks volunteers who are willing to be a part of the Translate Community and help improve Google Translates accuracy. There are two ways in which these volunteers can contribute to the service. Firstly, they can review phrases translated by Google, and submit their corrections if they feel Translate has made a mistake. Second, they can view all possible translations for phrases and then click-select the translation they feel is most accurate.

9.3.5 MICROSOFT TRANSLATE

Introduction

The Microsoft Corporation has created several state-of-the-art technologies in the areas of natural language processing, speech recognition, dialog systems and spoken language understanding to help computers master the nuances and complexities of human communication. Microsoft Inc. started developing the first version of its machine translation system in 1999-2000 and has enormously developed its multilingual range and technology since then. Here, the focus shall be on Microsoft's work on Indian languages and explore some of its key innovations in the ICT sector, keeping in mind their implications for India.

Structure and Funding

As a private Software Developing company, Microsoft invests in language technology from its own capital to make its apps, Windows and Office compatible for multilingual solutions.

Research Areas

Deep Learning

It works towards advancing the state-of-art in deep learning developing algorithms, models, and systems in deep supervised and unsupervised learning, deep reinforcement learning, and neural-symbolic reasoning, and then pursue breakthroughs in natural language processing, computer vision, conversational AI, multimodal intelligence, and other relevant areas.

Dialog Systems

It aims towards developing language technology that helps people carry out live conversations and interactions on multiple platform.

Speech Recognition

A project called 'Spoken language understanding' is researching the emerging field in between the areas of speech processing and natural language processing. It covers tasks such as domain detection, intent determination and slot-filling, using data-driven methods.

Machine Reading

A large-scale conversational question-answering dataset made up of conversational questions on a set of articles from different domains is known as 'CoQA'. These questions are used to mimic human interactions and learn to interact with humans.

Key initiatives

Conversations in Real-time

Skype and Microsoft Translator have enabled people across the globe to have real-time conversations by facilitating simultaneous translations and subtitles for over 60 languages

Preserving Endangered Languages

It also allows organisations to create domain-specific systems, including industry-specific translation systems (for instance, for the medical or financial sectors) and business-specific systems that are customised to the company's internal style and terminology.

Localising software's

Microsoft Windows, Microsoft Office and other software can be powered to be used in the languages supported by Microsoft Translator. This makes them more attractive for business and personal use in countries where English is not the first language.

Presentation Translator for PowerPoint

This add-in for Microsoft PowerPoint provides subtitles for live presentation. Users can choose to translate and view presentations in any of the available languages.

Bhasha India:

Under its Bhasha India programme, Microsoft has developed a number of customised tools to create a work environment conducive to Indian users such as mentioned below

- The Microsoft Indic Language Input Tool (ILIT) which allows users to easily enter text in Indian languages into any Microsoft Windows application using transliteration as its primary input mechanism while providing a visual keyboard to edit the text that does not get transliterated properly. It supports 22 Indian languages.
- The Microsoft Captions Language Interface Pack (CLIP) uses tooltip captions to display translations for English user interface terms. CLIP is designed for those Visual Studio users who are not very fluent in English and can help them learn and use Visual Studio 2010 by providing translation for the most common user interface elements of the Visual Studio Integrated Development Environment (IDE). CLIP is the result of close collaboration between Microsoft and local academic communities. It is available for download in Hindi, Odiya, Tamil and Malayalam.
- The Multilingual App Toolkit (MAT) is an integrated Visual Studio tool that allows developers to streamline localisation workflows of their Windows, Windows Phone and desktop apps. MAT improves localisation of file management, translation support, and editing tools.
- The Microsoft Translator is a cloud-based automatic translation service that can be used to build applications, websites, and tools requiring multi-language support. Any of the languages supported by the service can be translated by using the Microsoft Translator Text API. Additionally, Microsoft Translator is integrated into Microsoft Speech Services;
- An end-to-end REST based API that can be used to build applications, tools, or any solution requiring multilingual speech translation. The Microsoft Translator can translate about 5000 characters at one time in 60+ languages.
- Speech-to-speech and speech-to-text services are available for all of the supported languages.
- The Microsoft Edge browser is capable of translating the web pages in foreign languages to any of the 60+ languages that it supports.

- Microsoft text-to-speech is a neural text-to-speech programme that generates audio nearly indistinguishable from recordings of people talking. Using AI, natural phone inflection and articulation can be added in to reduce listening fatigue. This technology can be harnessed to create audio books from e-books, enhance in-car navigation systems by making them interact with chat-bots and virtual assistants, etc.

9.4 GENERAL COMPARISON

After obtaining an understanding of various language technology initiatives around the world, a general comparison has been performed amongst them based on select parameters such as the year of inception, nature of users, network spread, services provided, management and funding.

Initiatives in Language Technologies	EU	CANADA	African Language Technology Initiative (ALT-I)	MICROSOFT TRANSLATOR	GOOGLE TRANSLATE	TDIL
TYPE	International Union	Public	NGO	Private	Private	Public
YEAR FOUNDED	2010	2016	2002	1999-2000	2006	1991
NO. OF TARGET LANGUAGES	24	60	3	60+	100+	23
SERVICES AVAILABLE TOBENEFICIARIES	Machine Translation, API, e-commerce, e-health, NLP	Unicode fonts for languages; e-learning content for major works of literature	Windows and Office in three Nigerian Languages; Unicode Fonts	Translation across various programmes of Office, Webpages; Custom Translator	Text-Text, Speech-Speech, Image to text, Document translation, Transliteration in 20 languages	OCR, OHWR, multilingual machine-based translation, e-learning tools, Unicode fonts. Phonetic Engines in 12 languages
Beneficiaries (paid/ non-paid)	28 Member States (paid)	Research Institutions, students, government, public	Research Institutions, students, government, public	Professionals, students, general users, Technology for Software Developers on paid	Professionals, students, general users, Technology for Software Developers on	Research Institutions, Software Developers, Students, general

				basis	paid basis	users, State and Central Government Departments, Tourism Industry
AGENCIES INVOLVED/ MANAGEMENT	LT-Innovate, GALA, META-NET, World Wide Web Consortium, EXPloiting Empirical appRoaches to Translation (EXPERT)	Digital technologies Research Centre, National Research Council of Canada; Government of Canada	Microsoft Corporation, International Development Research Centre, TIWA Systems Limited, Open Society Initiative for West Africa	Microsoft Corporation	Google Inc.	Ministry of Electronics and Information Technology (MeitY), Government of India; Centre for Development of Advanced Computing, Pune
FUNDING	European Union	Government of Canada	Bait Al-Hikmah, OSIWA, IDRC, ACALAN, Lagos State Development and Research Council	Microsoft Corporation	Google Inc.	Ministry of Electronics and Information Technology (MeitY), Government of India

KEY INITIATIVES	META-SHARE, META-FORUM, CLARIN VLO (Virtual Language Observatory) Cracker Project, Digital Single Market, FREME, European Language Resource Coordination	Pirurvik, Creating Online Indigenous Language Courses (COILC initiative)	Standardisation of major languages, e-learning courses for the languages, Uni-coding African Languages	Bhasha India, Live-Conversation, Conservation of Endangered Languages, Office & Windows access	Google Translator Toolkit, Translation Community, Google Text-to-Speech, Google Open Source	National Rollout Plan, Uni-coding Indian Languages, Integrated platform for Multilingual Machine Translation, Web-Standardisation Initiative
AREAS OF RESEARCH	Hybrid Translation, Machine Translation, Deep Learning, User-based content creation	Standardisation of Indigenous Languages, e-learning resources for literature of Indigenous Languages	Automatic speech recognition; Text to speech synthesis; Machine translation; Spelling checker; Automatic diacritic application; Localisation of software; Assistance to Universities	NLP, language technology to aid education, user-based content creation	Multilingual translation across text, speech, Image, Document, website	NLP, Machine Translation, Text-to-Speech, Image-to-text (OHWR, OCR), e-learning tools, Speech-to-text

9.5 STUDY OBSERVATIONS

- In terms of availability to the general public, Google and Microsoft, although for-profit companies, provide their services for free to the general internet user whereas for sharing the API for their respective software or services, they charge the interested company or organisation for using technology developed by them. European Union, although an international organisation, also invoices its clients for sharing the language technology but it develops its technology primarily for its proliferation into the market. ALT-I, is a non-governmental organisation and it is funded by organisations and private companies, but it does not have properly developed technologies for selling it in the market, its goal is primarily to induct the African language into the digital world. TDIL, amongst all, is by far the biggest and longest running project which is directed towards citizen empowerment and is run by the national government. The tools developed under TDIL are available free of cost for the Indian academic researchers and start-ups, while for the MSMEs and International Academic researchers at 10% of the cost, for Big Companies, MNCs and foreign entities, at 50 % of the cost. The cost refers to the cost of creation of that resource at "current rates. The academic researchers are supposed to sign a non-disclosure agreement with the CDAC through which the technologies will be made available. Through this arrangement, it has been ensured that start-ups and small businesses do not have to rely on foreign and expensive technology while developing technologies for citizens
- Google Translation Community incorporates inputs from citizens on the internet for most sought after enquiries or prominent sentences and phrases where they can use it to ensure accuracy for that language. On similar lines, Microsoft has tied up with academia to create corpus for language pair translation by getting data from the research work. ALT-I, also works with universities and students to get young researchers engaged in linguistics and language technology and also uses radio broadcasting to create corpus for speech and language recognition. TDIL, on similar lines, can also collaborate with academia to include tonality and multiple dialects for the corpus creation. This will play an important role for enhancing the technologies like Speech recognition. While GO TRANSLATE by Government of India takes help from people to translate government websites into local languages, it has not many active projects (as other global initiatives have) for corpus creation of different languages which will be required for developing tools of the next round of language technology like Speech-Speech translation or live translation.
- It has also been observed that TDIL and ALT-I have dealt extensively with the issue of Uni-coding and Standardising and catered to the challenge of illiteracy and lack of digital infrastructure and therefore the path has been very difficult. Both of these initiatives are directed towards the empowerment of citizens. For European Union, the focus lies within the purview of strengthening business ties within the European businesses and communities and for Microsoft and Google it is about facilitating the users of their software-related services to reach a wider user-base.

- Canada has developed the Unicode for its Indigenous languages because they were being neglected in competition with the other languages such as English. Microsoft has also initiated some community projects for languages like Kiswahili, Hmong and Otomi which works not only towards developing technologies for languages but also for promoting some of the endangered languages so that they are used more widely. TDIL should also take care of the tribal dialects and languages in areas like those of the north-east where the languages differ very rapidly with variation in the area to preserve the languages of the people regardless of the low population that is related to the language.

9.6 CONLUSION

TDIL has been doing a considerable amount of work to safeguard the precious language resources of India. Along the same line, several governments and private organisations around the world have been developing such technologies.

This chapter compared TDIL with some of the language technology initiatives running around globally such as Alt-I, European Union, Canadian Indigenous Language Project, Microsoft and Google. Through an extensive secondary research each of the initiatives were studied and compared.

It was identified that TDIL is one of the largest initiatives focussing on the challenge of illiteracy and lack of digital infrastructure. But as the case of other initiatives, TDIL needs to use feedbacks and surveys to re-centre its goals as it proceeds to develop technologies. This becomes more important since projects like these take time and the expectations and needs of the users could have changed from, they might have been. This gap might lead to corrosion of market share of TDIL as private corporations are quick to identify this gap and use it for making profits. Hence, it is high time that TDIL visualises its projects as being complimenting agencies to the market and start looking for 'utility' based approach to develop further if we want people to prefer the indigenous technologies over the ones developed by private corporations. Drawing inspiration from the Microsoft Edge browser, the indigenous languages must be provided with browsers in their own scripts because applications are limited in their scope while the browser enables one to look for information on a much larger scale.

It has also been observed that TDIL is still to have a repository or any ongoing collaborations for continuous corpus creation like ALT-I, Microsoft or Google which is a very important process for improving the quality of machine translation. State-sponsored universities as well as Central universities need to be empowered as many of them have distinct languages and well as developed corpus of literature to help with the development of corpus creation and e-learning tools.

Lastly, to complement the infrastructure yet developed in India to access the language technologies developed by TDIL, it is necessary that technologies are workable in offline mode that will make the tools accessible after they have been downloaded once.

Chapter 10: KEY RECOMMENDATIONS

Recognising the need and wants of the citizens, MeitY, Government of India, initiated the Technology Development for Indian Languages. The goal is to empower the citizens by enabling them to take on the various advantages of ICT available in India. Through the means of language processing tools various applications of the internet are now being made available to the masses of India.

TDIL programme since its inception has made significant impact in the field of natural language processing and digital world of India. After the analysis of the current phase of TDIL and understanding the need for the continuation of the programme, we have summarised some recommendations based on the five pillars of impact assessment, viz., Technical, Economic, Legal, Operational/Organisational and Social.

10.1 TECHNICAL

10.1.1 RE-THINKING AND RE-DESIGNING THE CORPUS

"There is an urgent need to develop a rich corpus, which is an essential building block in all language technology components, by facilitating data sharing between different government functionaries, organisations with TDIL."

Corpus developed for the TDIL programme so far has enabled to create a base for various language tools. However, in order to sustain the model and for its further advancement, the need is to re-think and re-design the module. Current corpus size of Indian languages is very

small for efficient development of Language Technologies in India, in comparison to other models of the world. India needs more extensive corpora which should be in billions, on the lines of its European counterparts. The existing team of experts in language technology can be collaborated to effectively enhance the corpora of Indian languages. Meanwhile some ideas are described below, for the same purpose.

Firstly, in light of the same, to build an efficient corpus the already available data with the government can also be utilised. As per the National Policy on Data Sharing and Accessibility (NPDSA) 2014, a chief data officer is designated in each organisation, ministry and department. These chief data officers should be given the responsibility of providing the data collected from different departments for the language development programme. A lot of the required data can be found on data.gov.in, the open data platform of the Government of India. Datasets are already available which can be used through web services on the portal.

There are a lot of departments the Gov. of India who have large amounts of archival data which can be directly utilised for training purpose. For example, to save time and not start the research from scratch, Doordarshan, widely known news and entertainment channel available at national and various regional levels, can provide with its dataset to build the corpus. Doordarshan as a government channel is an extremely rich source for building text and speech corpus for multiple Indian languages.

Secondly, departments and organisations should be incentivised for sharing the data. This shall moreover engage an active participation and strengthen the collaboration of various departments, thus, making a strong robust information hub within the nation. A standard process must be created by TDIL in order to identify which languages need to be prioritised for corpus building based on parameters like usage of that language by the people, impact on overall development of the community/region, etc.

10.1.2 DEPLOYMENT AND NEW ARCHITECTURE OF DC PLATFORM

"Deployment should be the next focus for TDIL and Language Tools should be developed in a 'Plug-and-play' kind of architecture. The next version of the DC platforms should be based on cloud infrastructure where Indian language infrastructure, services and components are available on the same."

While several prototypes have been developed by each of the consortia under TDIL, there is no operational machine translation system which can be used. On the other hand, there are many systems such as e-Translation in the EU which are being used by several organisations across the EU. Though a fully automated machine translation system still remains elusive, one can use the presently available MT systems involving a human in the loop. This is how other countries are using machine translation systems despite the fact that no perfect machine translator has yet been developed. The next phase of the programme must focus on the deployment of the systems.

In this context, it is advised to adopt two stage approach and plug-and-play architecture. First, a system should be developed using open source software for MT and freely available parallel text corpus. This should be made available to a selected group of users who are willing to use at this

stage. Once the MT systems are made ready by the consortia, these can be plugged into the platform. Similarly, when corpora or any tool is made available by the consortia, the same can be plugged into.

The obvious advantage of this approach is that some sort of the translation system would be available within a short span of time for those who are willing to use. This will create corpora over a period of time as all the translated material can always be used as corpora. At the same time, the platform can always get benefitted from the research being done by the consortia.

Language technology systems need to be interoperable and should have open APIs or other suitable methods to connect and communicate with each other. Components of language technology should be developed in a plug 'n' play kind of architecture where applications can be built by dragging and dropping the components. Ease of building and availability of resources will help in building the right environment for businesses to come out with new products and services. The start-up ecosystem is hungry to try out innovations and build products and services for this new set of non-English speaking consumers.

Next version of DC platform should be built on cloud infrastructure where Indian Language infrastructure, services and components are available on cloud infrastructure. Companies, organisations, departments, start-ups and individuals can use the configuration-based system to build the applications, platforms, products and services in Indian Languages. This will be consumed by the whole world as they see India as the biggest consumer market of the future.

10.2 ECONOMIC

Understanding the need to create a way forward for the TDIL initiative and analysing its further growth in the economic structure recommendations such as market creation and collaboration, thus, have been put forward as follows.

10.2.1 MARKET CREATION FOR LANGUAGE TECNOLOGY AND APPLICATIONS BY CENTRAL AND STATE GOVERNMENTS AND INDUSTRY

"There is a need to create space wherein individual language developers, translators or different organisations are set in place to enhance the research and development for language technology. While existing large companies and commercial players could also be convinced to provide products, services and platforms in Indian languages catering to non-English speaking population of India."

Modern language industry has developed rapidly following availability of the internet. Various tools include like that of translations of heavy texts within seconds without the need of human intervention. As a business model, it has allowed small scale industries to have international expansion and growth. The Slator Language Industry Market Report 2019[24] provides a

[24] https://slator.com/data-research/slator-2019-language-industry-market-report/

comprehensive view of the global language services and technology industry, which, according to Slator, was a USD 23.2 bn market in 2018 and projected to grow to USD 28.2 bn by 2022.

Central and State Governments should help in creating jobs and business opportunities in the applications of Indian language technology. Various central and state government departments are still using manual translation and other language services instead of using tools and organisations providing these services. Small language technology companies are not able to survive because of lack of opportunities from the departments. A huge size of digitisation and translation work is available with the government departments which should be distributed with the language technology companies.

Lastly, existing big companies and commercial players could also be convinced to provide products, services and platforms in Indian Languages catering to non-English speaking population of India. For example, e-commerce websites rendering their service of online shopping in Indian languages, or banks providing the online and offline banking and payments services in Indian languages can bring the kind of transformation that is required in Indian language usage industry.

10.2.2 BUILDING UP A START-UP ECOSYSTEM

"Start-ups can play a beneficial role in the commercialisation process, getting more power and innovative ideas to digitally empower the citizens while attracting the global investment market."

Start-ups providing different technology services and applications in Indian Languages need to be promoted by the ministry. Seed funding should be provided along with the incubation support, marketing exposure and the technology tools to flourish the business environment.

India currently is a growing hub for various start-ups and sees over 3100 start-ups a year. This sphere of tech start-ups, if continues at the same pace will be able to generate over 2.5 lakh jobs in the next 5 years. It is because these endeavours in the field of language processing and technology are beneficial when seen in the context of the commercialisation process. As they work to involve more manpower and innovative ideas that eventually work towards the digital empowerment of the people. A lot of work has already been put in place by various researchers and developers to produce digital technology in local Indian Languages. Additionally, a favourable start-up ecosystem is attractive in the global market, as the growing number of start-ups work to lure foreign investors to the country and make India an ideal destination for foreign investment. In recent years, there have been several initiatives by other countries to link up their businesses with Indian start-ups. This further highlights the need and importance of more power and investment into these initiatives depending upon their role in the growth of India.

10.2.3 EFFICIENT AND CONTINUOUS FUNDING OF THE PROGRAMME

"A continuous and systematic funding procedure needs to be put in place for the TDIL programme."

Funding is another critical requirement of the TDIL project. A lot of activity is happening on developing the corpus, tools and collaborations both on domestic fronts and with a scope of

global collaborations. To further streamline the processes of developing our research and development methods and tools, there is a need of efficient funding for the project. The funds thus also need to be released timely by the respective authorities.

The outcome of the survey results and understanding of different stakeholder perspective points to the need of sufficient and regular funding of TDIL projects. R & D, deployment and commercialisation efforts cannot be made effective without a well though and planned funding in place.

10.3 LEGAL

Based on the legal framework and analysis of TDIL the further section comments on the licensing of various tools developed under the programme.

10.3.1 OPEN SOURCE AND FREE LICENSING

"The technologies and components developed under TDIL should open source and available free of cost to users and development community."

Firstly, there is a need to modify the current guidelines available for the research and development, especially focusing on the Intellectual Property Rights. Currently, the IPR is with the institutes and the commercialisation and proliferation efforts are not creating much results. The technologies created under the initiative should be available under the General Public License (GPL) for further use and development. The open source community can develop the technology and build multiple solutions for the industry and users. Another advantage of open source is that the other technology providers also build connectors, etc., for mutually beneficial purpose and that helps in making systems scalable and interoperable.

Currently, TDIL is building a new portal to monetarily capitalise on the resources of the language technology. The resources should be provided free of cost or at a very low price. The main or primary focus should be on building a strong and effective ecosystem, as it will, in return, produce more outcomes in comparison to the sale of translation tools. This will result in an increase in literacy rate, better healthcare services and participation, as well as in-flow of cash in banks, thus, further enhancing the flow of money in the country itself. Such goals and aims should be more focused upon as they will lead the way to a more sustainable development.

10.4 OPERATIONAL/ORGANISATIONAL

TDIL as any other initiative requires a strong and robust structural model of implementation to give out best possible outcomes. Keeping that in mind several operational and organisational recommendations have been put forward for the programme.

10.4.1 DICTATING THE VISION AND MISSION

"Primary focus should be shifted towards defining the vision of the initiative with detailed and realistic goals as well as deliverables."

Firstly, in order to initiate a programme, there is need to define set goals and aspirations for the project under a well-defined, coherent, hierarchical structure which, by its virtue, should empower individual projects through adequate focus and allocation of resources. The aim here is to redefine the vision statement of the TDIL project by making more realistic approach. While the vision of TDIL is *"Digital unite and Knowledge for all"* and the mission is *"Communicating & moving up the knowledge chain by overcoming language barrier",* this can be further sub-divided into mini practical goals which are more visible and deliverables that lead to applications. In order to achieve these targets, firstly an understanding needs to be made that TDIL programme is not a solution, but rather a way forward to a sustainable development. Secondly, a roadmap needs to be created with a set number of tasks in hand pertaining to each phase of development. Lastly, creation of timeline should be done to, one, overcome the current problems that have arisen by taking up large number of tasks at one-go, and two, will help in defining the goals for better efficient outcome.

10.4.2 CONTENT CREATION

"There is an urgent need to expand the amount of content available online in Indian languages."

The amount of online content available in Indian languages is drastically low and this also prohibits the masses to use internet or ICT. Applications of ICT worldwide are majorly available in English. Therefore, various independent organisations and researchers create content which is limited to English only. Central and state governments, organisations and departments should aim to create independent content in Indian languages for various domains and make it available online on different platforms. The central and state governments need to outsource the creation of content to begin with, so as to expand the information available online in Indian languages.

10.4.3 PROJECT MANAGEMENT UNIT

"To operate the initiative in an enterprise project management mode there is a need to establish an effective Project Management Unit (PMU)."

TDIL as an initiative can be seen as an independent project as it, like any other project, requires a strong framework for it to be well executed and deliver its best outcomes. This framework further facilitates in organising the enterprise's resources so that there is a direct relationship between the leadership's vision, mission, strategy, goals, and objectives to move the organisation forward. The next step calls for continuous monitoring with alerts and notifications so that management has granular visibility of task and is able to see the green, yellow and red flags against the project. Effective management can also be implemented through operational dashboards linked with accountability metrics and a risk management plan. An accountability plan can be an important feature for the entire life cycle of the TDIL programme as it could be used at different stages to keep a check on its progress. An effective management, thus, aims to provide best solutions and helps the programmers to keep a track record of their progress both in terms of growth and fall, also allowing them to make effective changes to them at the same time.

10.4.4 GOVERNMENT AND E-GOVERNANCE SITES

"Initially Government and e-Governance sites should be translated in Indian local languages."

Language processing and translation of Indian Languages should be first implemented on government sites and e-governance services and mobile apps to start with. For instance, government apps like UMANG could be enhanced and developed further, so that it can render services in all 22 officially recognised languages. This will directly impact the availability and accessibility of information and services to the larger masses. Secondly, in light of the same, Data Centre Portal could be made available in all TDIL supported languages, at least to start with.

10.4.5 STATE LANGUAGE MISSIONS

"State IT officials and secretaries should be invited on board."

To build the TDIL corpus more efficient several state IT secretaries should be invited on board. This can be further implemented in two different steps. One, they should be sensitised and pave the way forward into leading to the creation of a unit in their respective state's IT department or the selected state language body. The unit's major focus should be on creation of linguistic resources, associated standards and basic tools pertaining to state languages. Furthermore, specialised training and awareness campaigns should be conducted for officials along with focusing on deployment of language technologies in state e-government services in collaboration with participatory start-ups. These units should work closely with the Government of India and its efforts in this direction as proposed in National Translation Mission.

Two, sub language projects/missions for various individual languages could be set in place. The ownership for the same will be given to the state governments respectively. This shall help in development of corpus and further facilitate the participation between the centre and the state.

10.4.6 CONTINUOUS EVALUATION AND ASSESSMENT WITH INTERNATIONAL STANDARDS

"Continuous evaluation and assessments can be put forward tabling the deliverables and outcomes of TDIL in comparison to those available in the international market."

TDIL has been doing technical evaluation of its projects of R&D and deployment but it does not give the wholesome status of the programme. A continuous evaluation and assessment framework is required for the TDIL to assess its progress against the vision. The evaluation should be done on an annual basis and should be user and need based. This framework should follow the global standards for programme evaluation. The learning from the similar initiative study can also be put into the framework.

10.4.7 AWARENESS, BRANDING AND PUBLIC RELATIONS UNIT

"There is a need to make people aware and informed about these initiatives so that they are able to make use of these facilities."

The TDIL programme will not only have an impact on the digital advancement of India but will also have a far greater impact on the social and economic growth of the country. This is, however, possible only if the people are made aware and informed about these initiatives. To ensure this, a special unit could be setup which will work towards connecting and educating people about the various services offered by the TDIL programme. This could be facilitated

through social media marketing experts who closely study the needs and grievances of people to provide them with the best possible solutions for the same.

10.4.8 SUPPORT SYSTEM WITH ESCALATION MECHANISM

"There is a compelling need to establish a comprehensive support system as well as an escalation mechanism."

As the number of projects under TDIL increases, so will the number of resources to be used by developers for creating end-user products and also the number of end-user applications in the market. Hence, there is a need of a comprehensive support system to be put in place for both the citizens as end-users and the developers for technical support and operational problems. Dedicated support hotline numbers can be put in place where developers can connect to relevant professionals from the institute whose resource they are using as a foundation. Similarly, for citizens, online and offline help centres and channels must be created to ensure proper product adoption and proliferation. These channels can range from customer service phone numbers to sms/whatsapp chat helpdesks to offline kiosks. In case of non-resolution of issues, an escalation matrix needs to be designed and effectively set in place to deal with potential problems, such as, for developing individual tools, assisting the public with the working and functions of the tools, commercialisation support, start-up support, etc. The Turn-Around-Time and quality of grievance resolution must be measured through user feedback to ensure effectiveness of the system and the grievances must be looped back to the partner R&D institutes for continuous technological improvement. An effective escalation plan shall deal with potential problems, such as, for developing individual tools, assisting the public with the working and functions of the tools, commercialisation support, start-up support, etc.

10.5 SOCIAL

Lastly, based on the need of the social parameters and how it influences the citizens. Certain schemes have been proposed under this heading.

10.5.1 DEVELOPMENT OF HUMAN RESOURCE

"An environment of learning is needed exceedingly for the proliferation of Indian Language Technology usage in masses."

Firstly, training programmes and workshops should be organised at frequent intervals for the officers and torch bearers in different geographies. These programmes should be extended not only for the technology and tools usage but also for further implementation of these services along with the commercialisation process and marketing. This will build the right environment for the language technology consumption. Exposure to newer technologies and areas, viz., NLP, AI and Neural Networks, are especially required for understanding the new paradigms of developments in language technology. Secondly, online programmes can be created by the TDIL in partnership with some of the training development organisations like C-DAC. These programmes should be made available to all the stakeholders. This in turn shall help in distribution of knowledge, standardise methodology and less expenditure on instructor led training.

10.5.2 SECTOR WISE PRODUCTS AND SERVICES

"There is need to keep a multi sectoral approach in mind while developing tools and making them available in the market for the already available services in the local languages."

The intended development for any project should not only be with regards to creating and advancing the tools within it but should also be focused on giving it a purpose. This can be done by making the project user centric and applicable within various sectors of the society. The major aim for the TDIL programme currently should be on the social impact it has, wherein the already existing services are made available to the people in their local languages. It could also work towards enhancing the information and service environment by making the developed language tools and services available for different sectors in the country. In the field of education for instance, the TDIL programme can provide already developed tools to facilitate the translation of the study material from English to various local Indian languages. This will work towards engaging more people into learning and higher studies rather than dropping out of school after primary education. Tools like a screen reader and audio books can further enhance the participation of people, especially people with visual and physical challenges and help them claim their 'Right to Education.' In the healthcare sector, similarly, where Artificial Intelligence (AI) is already playing a huge part, the implementation of TDIL will even help people from weaker economic sections and rural areas become aware and informed about the various services intended for them in their own language. Farmers can make use of information in their local languages for weather forecasts, government schemes, policies, etc. The banking sector has always been at a slight disadvantage because of language constraint as most of their services are only available in Hindi or English. The introduction of the applications and services developed under the TDIL programme can help make services like Net Banking available to people who don't belong to an English-speaking background. It can also facilitate to improve the citizen engagement in the sector, as it will help remove the mistrust and hesitation that people face regarding the sector simply because they are not aware about the various banking policies and services that the place offers. This in turn will help increase the deposits in the bank further contributing to the money flow within the country thus creating a fertile ground for various e-commerce initiatives in the consumer market to flourish.

10.6 WAY FORWARD

Intellectual wealth is a matter of great pride for any country because the kind of power that comes with knowledge is long lasting and permanent in nature. Through TDIL, India is striving rapidly to be accessible for and connected with the citizens to implement effective e-Governance. Such a robust, scalable and inventive programme goes a long way to successfully contribute to the country's economy.

TDIL since its inception has established a foundation in the field of language technology for future developments. The computing power and the available technologies were not good enough for the language technology development and its requirements. However, the emergence of high-power computing and Artificial Intelligence has paved the way for effective machine translation, natural language process and language technology programmes.

For instance, online-portals like the Sandhan Portal which is a Cross Language Information Access for Indian Languages have empowered citizens towards a more self-reliant stage, wherein, language barrier is no longer a hurdle to overcome in order to use the applications of ICT. Similarly, better language processing tools and the information available through TDIL has helped India in strengthening its capability through Indian local languages in various sectors.

Under various TDIL projects Linguistics Resources like Text Corpus, Speech Corpus, Dictionaries, word net, Script Grammar, checkers, etc are being developed. Machine translation and other functional applications has claimed attention from some of the inquisitive minds in linguistics, philosophy, and computer science to reach out to the common man across various sections. With the development of such applications it seeks the attention of not only common man but the rural population as well.

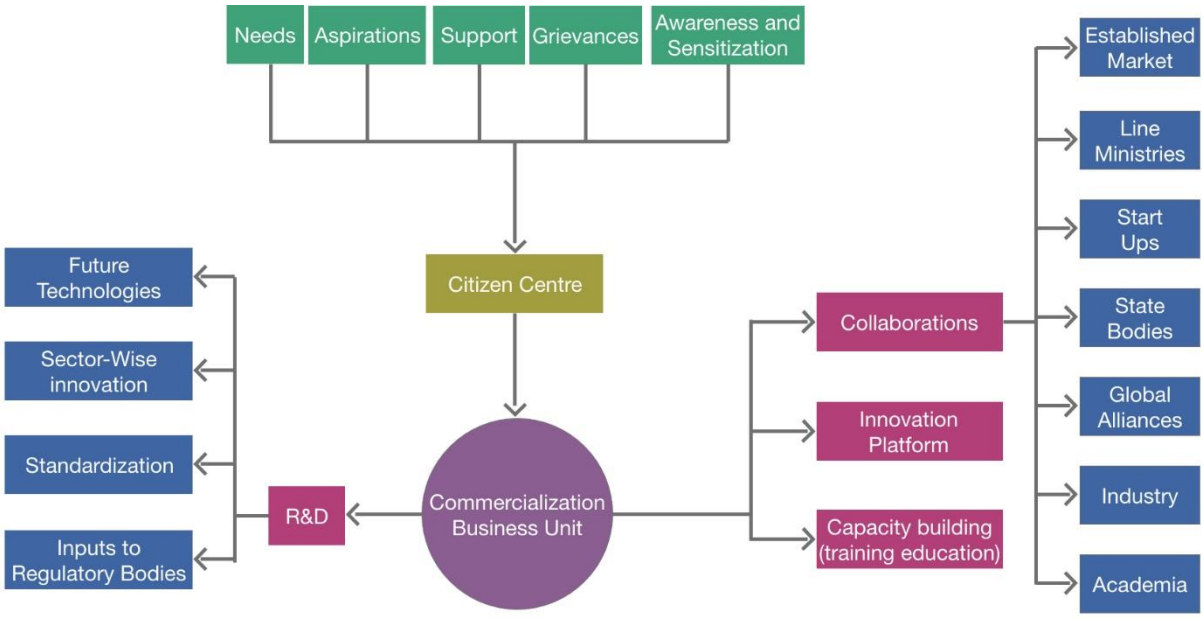


Figure 10. 1: Commecialisation Business Unit Architecture

Firstly, TDIL should set up a Commercialization Business Unit as a critical aspect and development for the programme. This business unit should be supported by a citizen centre that takes care of the support, grievances, needs and aspirations along with the awareness and sensitisation activities. This business unit gets the inputs from the citizen centre to plan the further activities based on need analysis. This information will further feed the R&D activities that redefine the core research areas by focussing on future technologies, sector-wise innovations and Standardisations. The R&D will also be able to provide the inputs to regulatory bodies to refine and strengthen the compliances. This business unit should have in its scope the ability to provide active collaborations between institutes, start-ups, industry, academia and Government both locally and globally. Availability of an innovation platform to complete the lab to land journey and empowerment of start-up ecosystem will be an added advantage of this business unit. As it is this business unit which has the complete information about the various needs and existing applications for Indian Language technology, it should also be the one looking after the complete building capacity for the technology right from education to training.

Secondly, the need is to bring about a in shift in the research methodology for an effective and efficient corpus building which will in turn facilitate the development of tools for sectoral growth

of the Indian economy. The sector wise innovations could contribute towards further development in the field of banking, agriculture, tourism, education etc. The tourism sector for example, has contributed 9.2% to the total GDP of the country in 2018, with the innovative technology of natural language and processing it can lead a way forward towards an expansion of tourism market. With the availability of means of communication through TDIL it shall further ease the delivery of products and outcomes. Thus, multiplying the results and directly impacting the economy. Similarly, in the education sector wherein various online lectures are readily available in English can thus be translated into local Indian languages to further enhance the flow of information and allowing citizens to learn and educate in one's own language. Banking services could be made easily available for the citizens who were earlier deprived of the services because of language barriers.

Thirdly, a proper set guidelines and standards need to be put in place. Along with standardizing bodies to create a unified standardization process of the tools developed.

Fourthly, it is important to take into account the emerging technology scenario. TDIL is about to start the next journey having a new set of expectations that will be more dynamic in nature due to rapid technology changes. It has a larger scope to be on the front run in adapting new and emerging technologies. TDIL should engage institutes and groups to have core capabilities in emerging technologies: Big Data, Artificial Intelligence (AI), Internet of Things (IoT), Robotic Process Automation, Hardware Robotics, Block Chain, etc. New products, services and platforms can be built on these upcoming technologies.

Although TDIL primarily comes across as a technology infrastructure initiative, its genesis and cultural mind set has always been centred around technology, whereas the fact stays that 'technology infrastructure' has been just one of the many aspects of TDIL's impact since its establishment. TDIL is in a position to reorganise itself as a 'knowledge empowered organisation' where technology infrastructure would be one of many enablers it would provide. Such a redefined holistic approach would help establish TDIL in its next phase as the empowering platform for enhancing the knowledge economy of the nation.

After this comprehensive and methodological evaluation, IIPA concludes that TDIL can become an indispensable part to the government and R&D and sectoral community for empowering the citizens. Observing the vital need of TDIL, IIPA lauds its potential and recommends enhancing its research, deployment and commercialisation processes along with the right funding by taking in account the future of this technology and empowerment that it can bring to the citizen.

Glossary

S No.	Abbreviation	Description
1.	ACALAN	African Academy of Languages is a Pan- African organisation founded in 2001 for the harmonization of Africa's many spoken languages.
2.	ACR	Abstract Character Repertoire is the set of characters to be encoded within the Unicode Character Encoding Model.
3.	AI	Artificial Intelligence is an area of computer science that emphasises on the creation of intelligent machines that work and react like humans.
4.	ALT-I	African Language Technology Initiative was launched with the aim of to appropriate various aspects of human language technology to facilitate human-human and human-machine communication in African languages.
5.	AnalGen	Analyze and General
6.	Angla MT	English Machine Translation System is a Rule Based Machine Translation System, designed for translating text in English to Indian Languages such Hindi, Punjabi, Bangla, Urdu, Malayalam and Telugu.
7.	API	Application programming interface is a code that allows two software programmes to communicate with each other.

8.	ASR	Automatic Speech Recognition is the use of computer hardware and software-based techniques to identify and process human voice.
9.	AU-CEG	Anna University College of Engineering is an engineering institute in Chennai, Tamil Nadu
10.	AU-KBC	Anna University KB Chandrashekar Research Centre in the Madras Institute of Technology in Chennai, Tamil Nadu
11.	BIS	Bureau of Indian Standards is the national Standards Body of India working under the aegis of Ministry of Consumer Affairs, Food & Public Distribution, Government of India
12.	BMP	Bitmap is a digital image composed of a matrix of dots
13.	CAG	Comptroller and Auditor General of India is an authority which audits all receipts and expenditure of the Government of India and the state governments, including those of bodies and authorities substantially financed by the government
14.	CBC	Canadian Broadcasting Corporation is a Canadian federal Crown corporation that serves as the national public broadcaster for both radio and television
15.	CCS	Coded Character Set is a character set in which each character corresponds to a unique number
16.	CD	Compact Disks can be used to store information which can be read by a computer
17.	C-DAC	Centre for Development for Advanced Computing is the premier R&D organisation of the Ministry of Electronics and Information Technology (MeitY) for carrying out R&D in IT, Electronics and associated areas
18.	CEF	A Character Encoding Form is the mapping of code points to code units to facilitate storage in a system that represents numbers as bit sequences of fixed length
19.	CERTIn	Indian Computer Emergency Response Team, designated to serve as the national agency to perform functions in the area of cyber security
20.	CES	A Character Encoding Scheme is the mapping of code units to a sequence of octets to facilitate storage on an octet-based file system or transmission over an octet-based network
21.	CHILDES	Child Language Data Exchange System is a corpus that serves as a central repository for first language acquisition data

22.	CIEFL	Centre for English and Foreign Language is a university for English and foreign languages located in Hyderabad, India
23.	CLDR	Common Locale Data Repository is a project of the Unicode Consortium to provide locale data in the XML format for use in computer applications
24.	CLIA	Cross Lingual Information Access
25.	CLIP	Microsoft Captions Language Interface Pack is a simple language translation solution that uses tooltip captions to display results
26.	CLIR	Cross Lingual Information Retrieval refers to the retrieval of documents that are in a language different from the one in which the query is expressed
27.	CoEs	A Centre of Excellence is a team of skilled knowledge workers whose mission is to provide the organisation they work for with best practices around a particular area of interest
28.	COILC	Creating Online Indigenous Language Courses
29.	CoQA	Conversational Question Answering systems is a large-scale dataset for building Conversational Question Answering systems
30.	CORPUS	Capable of Representing Potentially Unlimited Selection of Texts
31.	CPG	In Computational Paninian Grammar, a sentence is analyzed in terms of dependency relations, more specifically modifier-modified relations
32.	CRIM	Computer Research Institute of Montreal is an applied research and expertise centre in information technology, dedicated to making organisations more effective and competitive through the development of innovative technology
33.	CSA	Common Sense Advisory is an independent market research firm
34.	CSS	Cascading Style Sheets is a style sheet language used for describing the presentation of a document written in a markup language like HTML
35.	DC	Data Centre is a facility that centralises an organisation's IT operations and equipment, as well as where it stores, manages, and disseminates its data
36.	DC Portal	Data Central Portal

37.	DIT	Dehradun Institute of Technology is a best private engineering university in Dehradun
38.	DMZ	Demilitarized Zones is a physical or logical subnetwork that contains and exposes an organisation's external-facing services to an untrusted network, usually a larger network such as the internet
39.	DNCP	Dynamic Host Configuration Protocol is a network management protocol used on UDP/IP networks whereby a DHCP server dynamically assigns an IP address and other network configuration parameters to each device on a network so they can communicate with other IP networks
40.	DNS	Domain Name System is a hierarchical and decentralised naming system for computers, services, or other resources connected to the internet or a private network
41.	DoE	Department of Electronics
42.	DPI	Dots Per Inch is a measure of spatial printing, video or image scanner dot density, in particular the number of individual dots that can be placed in a line within the span of 1 inch
43.	DR	Disaster Recovery is an area of security planning that aims to protect an organisation from the effects of significant negative events
44.	DRM	Disaster Recovery Management
45.	DSM	Digital Single Market is a policy belonging to the European Single Market that covers digital marketing, E-commerce and telecommunications
46.	EBMT	Example Based Machine Translation is a method of machine translation characterized by its use of a bilingual corpus with parallel texts as its main knowledge base at run time.
47.	E-Gov	Electronic Governance is the integration of Information and Communication technology for delivering government services, exchange of information communication transactions, integration of various systems and stand-alone services between government to citizen, government to business, government to government, government to government employees as well as other processes within the government.
48.	EILMT	English to Indian Language Machine Translation System was developed under the TDIL Programme allows the translation of English to eight Indian languages, namely, Hindi, Bengali, Marathi, Urdu, Tamil,

		Oriya, Gujarati and Bodo.
49.	ELRA	European Language Resources Association is a non-profit organisation that was established to serve as a channel for the distribution of speech, written and terminology Language Resources for Human Language Technology.
50.	E-mail	Electronic Mail
51.	FAQS	Frequently Asked Questions
52.	FESTVOX	Festival Speech Synthesis Systems is a free software multilingual speech synthesis workbench that runs on multiple platforms offering black box text-to-speech, as well as an open architecture for research in speech synthesis.
53.	GALA	Global and Localisation Association is a global, non-profit trade association for the translation and localisation industry.
54.	GDP	Gross Domestic Product represents the total monetary value of all final goods and services produced within the country during a period of time. It is the most commonly used measure of economic activity.
55.	GIST	Graphic and Intelligence based Script Technology was developed by the Centre for Development of Advanced Computing, India for research and development in Indian Language Computing.
56.	GPL	General Public License is a widely used free software license, which guarantees the end users the freedom to run, study, share and modify the software.
57.	GSM	Global System for Mobile Communications is a second-generation digital mobile telephone network using a variation of Time Division Multiple Access. It is the most widely used digital wireless telephone technologies.
58.	GUI	Graphical User Interface is a visual way of interacting with a computer using items such as windows, icons, and menus, used by most modern operating systems.
59.	HLT	Human Language Technology studies the methods of how computer programmes or electronic devices can analyze, produce, modify or respond to human texts and speech.
60.	HTML	Hyper Text Markup Language is a standardised system for tagging text files to achieve font, colour, graphic, and hyperlink effects on World Wide Web pages.

61.	HTR	Handwritten Text Recognition is the ability of the computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch screens and other devices.
62.	IBM	International Business Machines Corporation is a global technology company that provides hardware, software, cloud-based services and cognitive computing.
63.	ICT	Information and Communication Technology is an extensional term for Information Technology that stresses the role of unified communications and the integration of telecommunications and computers as well as necessary enterprise software, middleware, storage and audio-visual systems, that enable users to access, store, transmit, and manipulate information.
64.	IDE	Integrated Development Environment is a software suite that consolidates basic tools required to write and test software.
65.	IDRC	International Development Research Centre is a Canadian federal Crown Corporation that invests in knowledge, innovation, and solutions to improve lives and livelihoods in the developing world.
66.	IIIT	Indian Institute of Information Technology is a group of educational centres in India that focus on information technology and associated business requirements.
67.	IISc	Indian Institute of Science is a public institute and deemed university for research and higher education in science and engineering located in Bangalore, Karnataka.
68.	IIT	Indian Institutes of Technology are autonomous public institutes of higher education in science, engineering and technology, located in India.
69.	IJCNLP	International Joint Conference on Natural Language Processing is an annual conference to discuss and learn about the various research and development in the field of Natural Language Processing.
70.	ILIT	Indic Language Input Tool helps a person enter Indian Language text easily into any Microsoft Windows application.
71.	ILT	Instructor Led Training is the practice of training and learning material between an instructor and learners, either individuals or groups.
72.	ILTM	Indian to Indian Language Machine Translation is a multipart machine translation system developed with the combined efforts of 11 institutions under the TDIL programme that has created language

		technology for 9 languages that have resulted in 18 language pairs.
73.	ILTPDC	Indian Language and Technology Proliferation Deployment Centre acts as a repository for all the research work carried out by premier research institutes in India in the area of LanguageTechnology.
74.	INSCRIPT	Indian Script
75.	IPS	Intrusive Prevention System is a system that monitors a network o0f malicious activities such as security threats or policy violations.
76.	IR	Information Retrieval is the process of obtaining information system resources that are relevant to an information need from a collection of those resources.
77.	ISCII	Indian Script Code for Information Interchange is a coding scheme for representing various writing systems in India.
78.	ISI	Indian Statistical Institute is an academic institute of national importance as recognised by a 1959 act of the Indian Parliament.
79.	ISO	International Standards Organisation is an international standard setting body composed of representatives from various national standards organisations.
80.	IT	Information Technology is the study of systems, especially computers and telecommunications for storing, retrieving, and sending information.
81.	JNU	Jawaharlal Nehru University is a public University located in New Delhi.
82.	KBCS	Knowledge Based Computer Systems programme 1986 was the compilation of the initial efforts in India under the area of Machine Translation.
83.	LAMP	Linux, Apache, MySQL, and PHP
84.	LPMS	Localisation Project Management System is a web-based portal that provides high end system for daily localisation to small time localisation companies.
85.	LR	Language Resources refers to a set of speech or language data and descriptions in machine readable form, used for building, improving or evaluating natural language and speech algorithms, or systems, or, as core resources for software localisation and language services industries, for language studies, electronic publishing, international transactions, subject area specialists and end users.

86.	LSP	Language Specifications in computing is a documentation artefact that defines a programming language artefact that defines a programming language so that users and implementers can agree on what programmes on that language mean.
87.	LT	Language Technology studies methods of how computer programmes or electronic devices can analyse, produce, modify or respond to human texts and speech.
88.	LUNs	Logic Unit Numbers is a unique identifier for designating an individual or collection of physical or virtual storage devices that execute input/output (I/O) commands with a host computer, as defined by the Small System Computer Interface (SCSI) standard.
89.	MAT	Multiple App ToolKit works with Visual Studio to streamline the localisation workflow for Windows Store, Windows Phone and desktop apps.
90.	MB	Megabyte is a data measurement unit applied to computer or media storage.
91.	MCIT	Ministry of Communications and Information Technology
92.	MietY	Ministry of Electronics and Information Technology is an executive agency of the Union Government of the Republic of India. It was formed after the bifurcation of Department of Electronics and Information Technology.
93.	MT	Machine Translation is a subfield of computational linguistics that investigates the use of software to translate text or speech from one language to another.
94.	NCPUL	National Council for Promotion of Urdu Language
95.	NDSAP	National Data Sharing and Accessibility Policy is a policy by Government of India with the objective to facilitate access to Government of India owned shareable data and information in both human readable and machine-readable forms.
96.	NER	Named Entity Recogniser seeks to locate and classify named entities in text into predefined categories such as the names of persons, organisations, locations, expressions of times, quantities, monetary values, percentages, etc.
97.	NGO	A Non-governmental Organisation is a non-profit organisation which works towards the betterment of society.

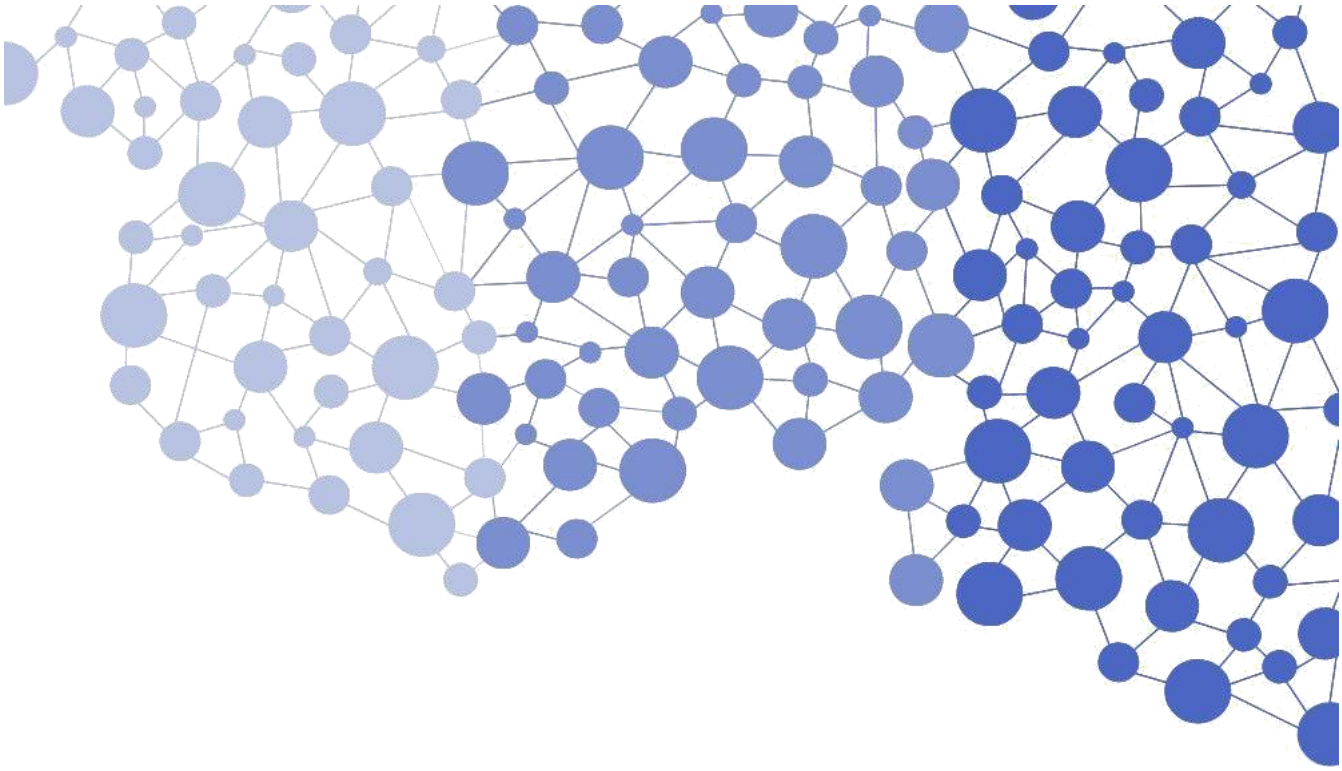
98.	NIC	National Informatics Centre is the premier science and technology organisation of the Government of India in informatics services and information and communication technology applications.
99.	NLP	Natural Language Processing is a field in machine learning with the ability of a computer to understand, analyze, manipulate, and potentially generate human language.
100.	NLP	Neuro-linguistic Programming is all about bringing about changes in perception, responsible communication and developing choices of responses or communication in a given situation.
101.	NLTM	Natural Language Translation Mission aims to make science and technology accessible to all by facilitating access to teaching and researching material bilingually — in English and in one's native Indian language.
102.	NMT	Neural Machine Translation is an approach to machine translation that uses an artificial neural network to predict the likelihood of a sequence of words, typically modelling entire sentences in a single integrated model.
103.	NPTEL	National Programme on Technology Enhanced Learning launched jointly by the Indian Institutes of Technology and the Indian Institute of Science and funded by the Ministry of Human Resource Development, Government of India is an online curriculum development programme in sciences and engineering at university and research levels.
104.	NRC	The National Research Council Canada is Canada's largest federal research and development organisation.
105.	NTP	Network Time Protocol is a networking protocol for clock synchronisation between computer systems over packet-switched, variable-latency data networks.
106.	OCR	Optical Character Recognition is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.
107.	OGD	Open Government Data is a philosophy- and increasingly a set of policies - that promotes transparency, accountability and value creation by making government data available to all.
108.	OHWR	In Online Handwriting Recognition System, the handwriting is captured and stored in digital form via different means. Usually, a special pen is

		used in conjunction with an electronic surface.
109.	OSIWA	Open Society Initiative for West Africa is a West African organisation which promotes democratic values.
110.	OWASP	Open Web Application Security Project is an online community that produces freely available articles, methodologies, documentation, tools, and technologies in the field of web application security.
111.	P3P	Platform for Privacy Preferences Project is an obsolete protocol allowing websites to declare their intended use of information they collect about web browser users.
112.	PMG	Pagemaker Group is the group formed by combining several objects to treat them with the help of a software application that enables individuals and groups to create and edit publications.
113.	PM-STIAC	Prime Minister's Science, Technology and Innovation Advisory Council is an overarching Council that facilitates the PSA's Office to assess the status in specific science and technology domains, comprehend challenges in hand, formulate specific interventions, develop a futuristic roadmap and advise the Prime Minister accordingly.
114.	R&D	Research and Development refers to innovative activities undertaken by corporations or governments in developing new services or products or improving existing services or products.
115.	REST	Representational State Transfer is a software architectural style that defines a set of constraints to be used for creating Web services.
116.	RPO	Recovery Point Objective is the age of files that must be recovered from backup storage for normal operations to resume if a computer, system, or network goes down as a result of a hardware, programme, or communications failure.
117.	RTI	Right to Information is an act of the Parliament of India to provide for setting out the practical regime of the right to information for citizens.
118.	RTO	Recovery Time Objective is the maximum desired length of time allowed between an unexpected failure or disaster and the resumption of normal operations and service levels.
119.	SaaS	Software as a service is a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted.
120.	SAN	Storage Area Network is a computer network which provides access to

		consolidated, block-level data storage.
121.	SHMT	Sanskrit Hindi Machine Translation System is an application whose objective is to develop Sanskrit to Hindi machine translation technology, using Sanskrit computational tools.
122.	SIGs	Special Interest Group is a community within a larger organisation with a shared interest in advancing a specific area of knowledge, learning or technology.
123.	SL	Source Language is the language being translated from.
124.	SME	Small and Medium-sized Enterprises are non-subsidiary, independent firms which employ fewer than a given number of employees.
125.	SMIL	Synchronised Multimedia Integration Language is a World Wide Web Consortium recommended Extensible Markup Language markup language to describe multimedia presentations.
126.	SMS	Short Messaging Service is a text messaging service component of most telephone, internet, and mobile device systems.
127.	SMT	Statistical Machine Translation s a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora.
128.	SNLP	In a Speech and Natural Language Processing Lab work is undertaken in many different sub-areas of NLP including syntax and parsing, semantics and word sense disambiguation, discourse and tree banking, machine translation, etc.
129.	SSF	Shakti Standard Format is a highly readable representation for storing language analysis. It is designed to be used as a common format or common representation on which all modules of a system operate.
130.	SSMT	Speech to Speech Machine Translation is a process that takes the conversational speech phrase in one language as an input and translated speech phrases in another language as the output.
131.	SWOT	SWOT analysis (Strength, Weakness, Opportunities and Threats) is a framework for identifying and analyzing the internal and external factors that can have an impact on the viability of a project, product, place or person.
132.	SYSLOG	System Logging Protocol is a standard protocol used to send system log or event messages to a specific server, called a syslog server.

133.	TACACS	Terminal Access Controller System refers to a family of related protocols handling remote authentication and related services for networked access control through a centralised server.
134.	TAG	Tree Adjoining Grammar in which the elementary unit of rewriting is the tree rather than the symbol.
135.	TDIL	Technology Development for Indian Language is an initiative to develop information-processing tools & technologies to facilitate human machine interaction in Indian languages and also to develop technologies to create & access multilingual knowledge resources.
136.	TELOS	Technical, Economic, Legal, Operational and Social
137.	TIFF	Tagged Image File Format is a computer file format for storing raster graphics images, popular among graphic artists, the publishing industry, and photographers.
138.	TL	Target Language is the language being translated to.
139.	TTS	Text to Speech is the technology that powers applications to read aloud the text on the screen with support for many languages.
140.	UNL	Universal Networking Language is a declarative formal language specifically designed to represent semantic data extracted from natural language texts.
141.	UTF-8	Unicode Transformation Format is a character encoding format which is able to encode all of the possible character code points in Unicode.
142.	UTM	Urchin Tracking Module parameters are five variants of URL parameters used by marketers to track the effectiveness of online marketing campaigns across traffic sources and publishing media.
143.	VAPT	Vulnerability Assessment and Penetration Testing are two types of vulnerability testing that perform two different tasks, usually with different results, within the same area of focus.
144.	VIVA	Vernacular Intelligent Voice Assistant provides support for voice-based interactions with the end-user using their vernacular languages.
145.	VLO	Virtual Language Observatory provides a means of exploring language resources and tools.
146.	W3C	The World Wide Web Consortium is an international community where Member organisations, a full-time staff, and the public work together to develop Web standards.

147.	WOFF	Web Open Font Format is a font format for use in web pages.
148.	WSD	Word Sense Disambiguation is an open problem concerned with identifying which sense of a word is used in a sentence.
149.	WSI	Web Standardisation Initiative is an initiative with the objective to promote and adoption of various internationalisation web and internet recommendations/ Best practices among developers, application builders, and standards setters, and to encourage inclusion of stakeholder organisations in the creation of future recommendations
150.	WWW	The World Wide Web is a combination of all resources and users on the internet that are using the Hypertext Transfer Protocol (HTTP).
151.	XHTML	Extensible HyperText Markup Language is part of the family of XML markup languages. It mirrors or extended versions of the widely used HyperText Markup Language, the language in which Web pages are formulated.
152.	XML	Extensible Markup Language is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.
153.	ZWJ	Zero Width Joiner is a non-printing character used in the computerized typesetting of some complex scripts such as the Arabic script or any Indic script.
154.	ZWNJ	Zero Width Non- Joiner is a non-printing character used in the computerization of writing systems that make use of ligatures.



Printed and Published by

Indian Institute of Public Administration (IIPA) Indraprastha Estate, Ring Road, New Delhi-110002. Fax.(O) +91-11-23702440, +91-11-23356528 E-mail: contact_us@iipa.org.in