Language Translation as a Socio-Technical System: Case-Studies of Mixed-Initiative Interactions

Sebastin Santy
t-sesan@microsoft.com
Microsoft Research
Bangalore, India

Kalika Bali
kalikab@microsoft.com
Microsoft Research
Bangalore, India

Monojit Choudhury
monojit@microsoft.com
Microsoft Research
Bangalore, India

Sandipan Dandapat
Microsoft R&D
Hyderabad, India
sadandap@microsoft.com

Tanuja Ganu
Microsoft Research
Bangalore, India
tanuja.ganu@microsoft.com

Anurag Shukla
Microsoft Research
Bangalore, India
t-ashukla@microsoft.com

Jahanvi Shah
Microsoft Research
Bangalore, India
t-jahshah@microsoft.com

Vivek Seshadri
Microsoft Research
Bangalore, India
visesha@microsoft.com

ABSTRACT

Seamless access to information in a rapidly globalizing world demands for availability of information across, ideally all but at the least a large number of, languages. Machine translation has been proposed as a technological solution to this complex problem. However, despite seven decades of research, and recently seen rapid progress in the field - thanks to deep learning and availability of large data-sets, perfect machine translation across a large number of the world's languages still remains elusive. In fact, it is a distant and perhaps even an impossible goal. Erroneous translations, on the other hand, can be detrimental in critical situations such as talking to a law enforcement officer; or, they could potentially perpetuate social biases or stereotypes, for instance, by producing mis-gendered translations. In this work, we argue that language translation is inherently a socio-technical system, which has to be viewed, studied, and optimized for, as such. The need and context of translation, the socio-demographic factors behind the human translators as well as the consumers of the translated content affect the complexity of the translation system, as much as the accuracy of the technology and its interface. Through a series of case studies on mixed-initiative interaction based approach to translation, we bring out the various socio-technical factors and their complex interactions that one has to bear in mind while designing for the ideal human-machine translation systems. Through these observations, we make multiple recommendations which, at the core, suggest that "solving" translation in the real sense would require more coordinated efforts between the technical (NLP) and social communities (HCI + CSCW + DEV).

COMPASS '21, June 28-July 2, 2021, Virtual Event, Australia

ACM ISBN 978-1-4503-8453-7/21/06…$15.00
https://doi.org/10.1145/3460112.3471954

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**; **Natural language interfaces**; *Collaborative and social computing systems and tools*; *Empirical studies in HCI*.

KEYWORDS

mixed-initiative interaction, interactive translation, socio-technical systems, human-agent interaction, human-centered AI

ACM Reference Format:
Sebastin Santy, Kalika Bali, Monojit Choudhury, Sandipan Dandapat, Tanuja Ganu, Anurag Shukla, Jahanvi Shah, Vivek Seshadri. 2021. Language Translation as a Socio-Technical System: Case-Studies of Mixed-Initiative Interactions. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '21), June 28-July 2, 2021, Virtual Event, Australia.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3460112.3471954

1 INTRODUCTION

Translation is key to making content available to different language communities around the world. It attempts to solve one of the central problems of a world with diverse set of languages, namely Information Exchange [45]. Translation is a crucial requirement for international trade[13], diplomatic relations, tourism [103], law [94], and crisis situations [76]. Historically, translations have been carried out by expert or professional translators who are not only proficient bilingual speakers but are also trained on the specific nuances of conducting translations. Recently, great progress has been made in automating translations by learning rules/functions over existing data that has been manually translated by humans. This area of work, called Machine Translation (MT), was one of the first problems researched within the artificial intelligence community [43] and has seen a recent resurgence with the amount of data and compute available at hand. As a result, the current translation industry is shaped by both human and machine translation, with the latter frequently used to aid human translators by providing automated translation suggestions. There is an ecosystem of interdependence as even machines also require human assistance in

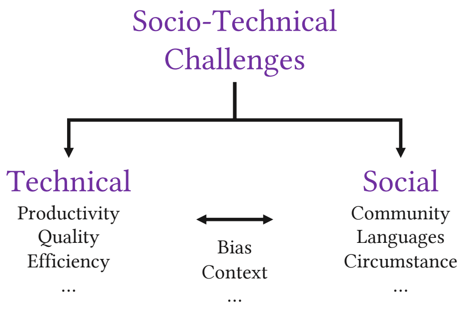


Figure 1: Socio-Technical Challenges of Machine Translation include both social and technical challenges as well as challenges that arise from the interaction between the two.

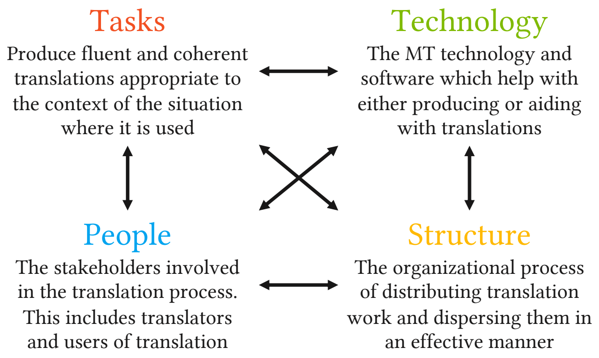

Tasks
Produce fluent and coherent translations appropriate to the context of the situation where it is used
Technology
The MT technology and software which help with either producing or aiding with translations
People
The stakeholders involved in the translation process. This includes translators and users of translation
Structure
The organizational process of distributing translation work and dispersing them in an effective manner

Figure 2: Language Translation as a Socio-Technical System where there is a continuous interaction between Tasks, People, Structure and Technology (Machine Translation).

form of vast volumes of manually translated data, which is critical for building state-of-the-art machine translation systems. Most of the research and practice as of now is aimed at improving translation quality while simultaneously increasing translator productivity and the overall efficiency of the translation process.

The task of translation while an inherently technical problem, has several societal implications [14, 94]. There are several social challenges which are required for the translation problem to be really "solved" (Figure 1). To highlight the issues, consider community-based translations [77]: the current (human and machine) translation setup is disproportionately favorable for languages with abundant data and expert translators. When combined with the fact that information is currently being churned out at a rapid rate, the current translation system is not scalable, especially since there are very few incentives to have such content translated at a worldwide level in a coordinated manner. This places such communities at a serious disadvantage, and at the opposite end of the spectrum where critical information is not sufficiently disseminated. However, such concerns are increasingly being addressed by community-driven efforts in which the incentives are frequently at a more personal or communal level rather than monetary ones. For example, in several of the community Q&A platforms such as StackOverflow or Quora, there are personal incentives of getting visibility as an expert in an area.

Whereas in the case of wiki platforms like Wikipedia, which mainly rely on anonymous contributions, the incentives are at a community level such as adding topics and notable people within their community and translating articles for their own languages [56]. Aside from these long-term efforts, there have been community-driven efforts in emergency crisis situations where information that is updated on a daily basis must be swiftly translated into several languages for widespread dissemination. In most situations, translation needs were routinely outsourced to bilingual speakers. Such setups have have shown to be effective in prior crisis situations such as during natural calamities like earthquakes on the Island of Hispaniola and Japan [15, 63] as well as during outbreaks such as CoVID [8] and Ebola [11]. While community-driven efforts are becoming more common in the space of translation, obtaining translations from amateur translators can have quality concerns and can be considerably slower to procure than their professional counterparts. Although community-driven efforts in translations rely on the amateur translators, there is a necessity for the transmission of reliable information as well as a need for a quick turnaround.

Thus, the process of translation is increasingly complex with a number of stakeholders in the process and hence merits to be looked at and solved as a socio-technical problem. In fact, as can be seen in Figure 2, translation consists of all the four pillars which constitute the framework of a socio-technical system [18, 61]. While fluent and coherent translations can be achieved through machine translation, context is difficult to capture as it is often *situated*. Machine-generated translations, for example, can be used for fairly benign tasks such as ordering food or in high-stakes situations such as interacting with law enforcement. Such a wide range of context variations necessitates varying levels of translation delivery (***Task***). The organizational framework for obtaining translations may also differ significantly. It might be a structured setup, such as through a language service provider (LSP), or it can be in a more democratized way, such as community translators working for crisis assistance and projects like Wikipedia (***Structure***). The process of translations has two primary stakeholders (i) translators, who carry out the translation, and (ii) users, who consume the translation. It is important to cater to the needs and expectations of each of these stakeholders (***People***). And finally, the underlying translation technology utilized can vary significantly, ranging from early rule-based systems through statistical MT and the most recent neural MT systems (***Technology***). There is no silver bullet that can help in navigating such a complex setup; however, previous approaches to such problems where humans and technology frequently interact have tried incorporating mixed-initiative techniques [42]. Mixed-initiative interaction refers to a flexible interaction strategy in which each agent (human or computer) contributes what it is best suited at the most appropriate time. Mixed-Initiative approaches have been much discussed in the field of translation [47]. Machines are good at recognizing patterns and have substantial amounts of memory and computing resources for large-scale processing. However, MT systems have not yet reached the levels of cognitive reasoning or pragmatic understanding to match that of humans [85]. Humans, though capable of performing these tasks efficiently in terms of reasoning and contextual understanding, can find them extremely laborious and tedious making such a setup expensive and tough to scale. Hence, there is a need for interactions to be designed to enhance the productivity of translators as well as optimize the quality of translations. While previous applications of mixed-initiative strategy have revolved around solving the technical issues in translation, we believe that it can be immensely helpful in tackling some of the social challenges as well.

In this paper, we investigate further into the role played by mixed-initiative approaches to help address some of the social challenges of machine translation. We explore this through the lens of different use cases which are set in varied contexts and have unique social challenges associated with them. These use-cases include enhancing the productivity of amateur translators, crowd-sourcing for translation data in low-resource languages, and incorporating visual context during translation and localization. All these three contexts are very different in the kind of community which participates or the stakeholders involved, the incentives provided as well as the output of the effort. We build mixed-initiative systems adapted to each of these use-cases individually and conduct preliminary need-finding interviews and pilot user studies to get feedback on them. It is important to understand how these deployments can affect each community and how can we improve on providing better interfaces for conducting translations in an effective manner. Through this work, we hope to make the community at large aware of how translation is a socio-technical problem that requires solving both technical and social challenges at multiple levels. Mixed-initiative translations have recently gained attention as it is repeatedly echoed within the translation community that MT cannot reach human parity on its own and would always require human-in-the-loop mechanisms [59].[1] Through the diverse set of systems we have built and preliminary studies we have conducted, we propose that in addition to aiding with the standard technical challenges which (machine) translation already faces, mixed-initiative translations can be an effective strategy at mitigating several of the social challenges as well.

The rest of the paper is organized as follows: §2 describes socio-technical systems and mixed-initiative translations in detail along with the previous efforts in the space. We also discuss the current capabilities and drawbacks of human and machine translation respectively to set the context for our work. This is followed by 3 sections: (§3 contains a detailed study which led to design explorations in §4 & §5) each of which covers a use case with a unique socio-technical scenario where mixed-initiative translation can help. For each use case, we discuss the challenges faced, the previous approaches used, our approach, and the feedback from the intended audience. Based on our observations and interviews, we provide some suggestions and recommendations in §6 which may help towards solving translation in a socio-technical sense and being inclusive of all the stakeholders involved in this process. We finally conclude the paper in §7.

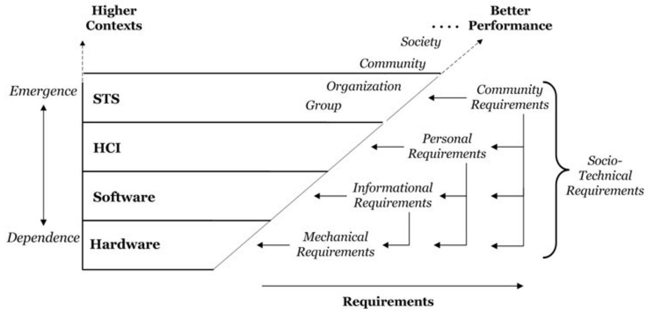


Figure 3: The different levels of a socio-technical system for computing, the dependence between those levels and the performance gain by incorporating each additional level into a system. [2]

2 BACKGROUND AND PREVIOUS WORK

2.1 Socio-Technical Systems

The notion of Socio-Technical System (STS) was introduced in the 1950s by the Tavistock Institute for manufacturing cases where the introduction of technology often collided with those of local communities, in their case, workers in English coal mines. This concept was later adopted into the computing literature to keep a check on the ethical use of computers [80]. When it comes to computing, the field started as a hardware problem and thereafter slowly including software which allowed for flexibility of interactions between different hardware. With the advent of the personal computing era, “people" started getting added to this equation so as to optimize for better user experience and interaction leading to the formation of Human-Computer Interaction (HCI) as a distinct discipline. As computing started pervading human lives and further affecting the community as a whole, it has essentially become an STS. To formally define STS in the context of computing, it is an isomorphic interplay between engineering, information, psychological, and sociological systems to achieve a greater objective (translation, in this case) [98]. Each of these systems is dependent on the other, and failure in any part of these individual systems can significantly impact the ecosystem and can lead to failure of reaching the desired goal [12]. Figure 3 shows how each of these systems is dependent on the other and with added requirements are able to incorporate better contexts and thereby impart better performance. However, although several of the computing systems such as software industries are considered STS, Ackerman [2] describes how there still exists a socio-technical gap between what computers do and what society wants [22]. STS have also been studied in the more fine-grained framework of "Web of System Performance" (WOSP) which constitutes of parameters like security, extendibility, reliability, flexibility, functionality, usability, connectivity, privacy [98]. For the scope of this work, we do not look at WOSP for translation.

Most discussed examples of STSs in computing include where computers are being used as a social medium aka social computing.

[1] Despite some success in achieving human parity in MT [41] which in general works in specific scenarios and test data.
[2] Figure taken from https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/socio-technical-system-design

Human Translation		Machine Translation	
Strengths	Limitations	Strengths	Limitations
Contextual and Pragmatic	**High Skilled Labor**	**Very fast, scalable and on-call**	**Literal Translations**
Humans are adept at reasoning and understanding the situation where a particular translation is required. For instance, differentiating between when translation is required for a informal context such as messaging vs. a formal context like writing an email or a book.	Human translators not only have to be bilingual speakers but should undergo appropriate translation related training in order to perform accurate translation. Often the professional translators use software to enhance their productivity and hence also require training for that.	Given the memory and computing resources which these translation systems possess, the translations produced by these systems are done so with minimal latency. Moreover, translations can be generated at scale with minimal supervision from humans.	Often by the virtue of automated translations being used in scenarios which require factual translations, they lack the capability to produce creative translations required for original work. This is mostly due to factual data being used to train such systems.
Knowledge Aware	**Slow and Unscalable**	**Several Languages**	**Massive Training Data**
Humans possess world knowledge and have real-life experiences at their disposal which helps greatly while carrying out translations. This can include awareness of proper nouns and other factual/world knowledge which makes the translation more robust and less prone to such errors.	As the translation process has several components ranging all the way from processing documents and assigning translators to ensuring quality of translations, the process is slow. For the millions of documents which need to translated even at the sake of quality, such a process is not easily scalable.	Once there are significant amount of parallel sentences in one direction, MT systems can be designed to produce translations to and fro from any combination of languages. This is unlike human translations which require fluent bilingual speakers to be able to translate.	MT systems in the current scenario require humongous amount of training data in order to generate fluent and coherent translations. NMT requires upwards of a 100k parallel sentences and SMT requires at least 10k sentences. This means that developing MT system is not feasible for most languages.
Metaphor Understanding	**Expensive**	**Diverse outputs**	**Biased and Non-explainable**
Humans are good at understanding certain complex nuances of languages such as idioms, puns, metaphors, lingos. Not only can these not be literally translated to another language, but doing so can even result in culturally non-relevant or possibly offensive translations.	Given that translation is conducted by highly-skilled translators, the process is often expensive. Translations are usually billed at per-word basis and thus can be costly to scale. Recently, this issue is addressed by crowd-sourcing translations which can be far cheaper but might comprise on quality.	Language is inherently divergent, and there are several ways of translating the same sentence. MT systems can easily enumerate all possible options to show the most plausible translations. Depending on the metric we are targeting such as adequacy/ or accuracy, different translations can be generated.	MT systems learn from real-world data which can be demographically skewed and hence can have biases which result in harmful or possibly offensive translations. This is further amplified by neural models which are hard to interpret and hence have constraints on how outputs can be modified.

Table 1: The strengths and limitations of Human and Machine Translation respectively. Most of these strengths and limitations between human and machine translation are complementary to each other, providing room for building better synergy between humans and machines.

These include emails, chatbots, e-commerce, blogs, wiki articles, social networks such as Facebook, Twitter, Youtube, and even search engines. When it comes to translations, they impact the daily livelihood of people and communities around the world and can often result in unintended consequences. Byrne [14] thoroughly discusses the consequences of translations in the legal, political and commercial contexts through multiple examples as well as talks about the translator's liability in these cases [33]. In a more positive light, as already mentioned, the advent of web technologies is allowing for community-level translations albeit with several social challenges associated with it [77].

2.2 Mixed-Initiative Translations

Human and machine translation have various strengths and weaknesses, and combining them can result in more efficient and high-quality translations. To accomplish so, it is necessary to understand what humans and computers are capable of individually, as well as how both may contribute to various elements of translation. [32]. Table 1 compares the strengths and limitations of human and machine translations. This allows us to see how an effective synergy may be formed between the two.

One way to achieve this synergy is to use interactive interfaces for translations, where machine translation techniques can assist with generating suggestions for human translators. Several efforts have been made to build interactive interfaces, primarily focused on two approaches to mixed-initiative translations: (i) post-editing or post-correction of translations, and (ii) directing MT systems using human input. Post-editing was one of the first ways of bridging human & machine translation and is currently the most widely used form of assisted translations available on translation softwares. Post-editing is the process where humans edit a machine-generated translation to the nearest acceptable form [5]. Although post-editing is the easiest form of interaction, the efficacy of this process is heavily debated [38, 39, 58, 60]. As language is inherently divergent, an alternate translation to the sentence instead of the suggested translation would necessitate considerable corrections, to the point where providing it would be useless. However, recent work has focused on the machine-in-the-loop approach (rather than human-in-the-loop in the case of post-editing), where the machine provides a supporting role to the process of translation. ITS [68] first proposed this mechanism due to the abysmal quality of machine translation at that time. Since then, there have been several tools developed in the context of translating webpages [30, 52] and independent systems such as Transtype [27, 57], CASMACAT [4], LILT [37], and Intelligo [23]. With the introduction of neural machine translation (Seq2Seq [10] and Transformers [93]), constrained decoding is a commonly used method for providing such suggestions [49, 84, 99, 106]. Several studies have been conducted to understand the efficacy of this approach [32, 37, 50, 66]. They suggest that interactive approaches such as post-editing and interactive suggestions reduces translation time, increases quality, primes the translator, and reduces drafting requirements significantly.

3 I: ENHANCING PRODUCTIVITY OF AMATEUR TRANSLATORS

Machine-aided Translation has long been an area of research due to the complementary qualities of human and machine translation. However, they have almost always been considered mostly from the aspect of increasing productivity for professional human translators. However, as community-based translations grow, considerable efforts are required to assist amateur translators by providing translation suggestions, especially in multiple languages. This can help translations to be produced at a faster rate, enabling for, for example, emergency information during a crisis to be swiftly disseminated across the community.

3.1 Background

Pratham Books is a non-profit organization that publishes affordable, quality books for children. It has published over 200 original titles in 280 languages and reached over 14 million children. In order to publish stories at this scale, and given that it is a non-profit initiative, they heavily rely on the community translators who are bilingual speakers without any required training as that of a professional translator. Pratham Books employs an interface called StoryWeaver[3] which helps in coordinating story-book writing and translations over the large set of books and languages. Once a storybook is written, community translators from different languages will take up the task of translating it to the languages they wish. These community translators include schoolteachers, community organizers, or individual benefactors working for the cause of their native language.

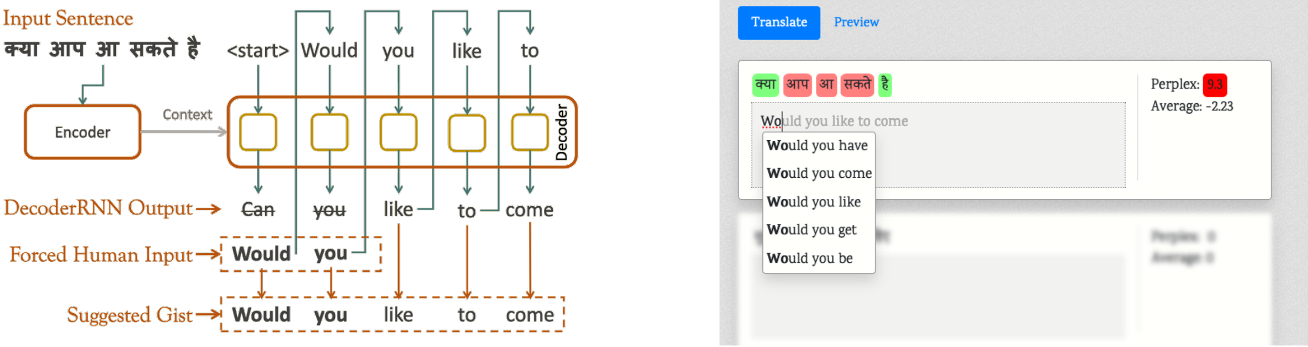
While using MT-generated translations straightaway is an obvious solution, it does not work in the intended manner. The key problems in this scenario are that (i) there are 280 languages that require to be translated, and the quality of MT systems quickly deteriorates after the top 10 languages, and (ii) that Pratham Books primarily has storybooks that require contextual and creative translation as opposed to literal translations generated by MT systems. However, these machine-generated translations can be provided as suggestions and cues which can help with boosting their productivity and quality margins significantly.

3.2 Challenges

To come up with a solution to this problem, we first identified the unique set of both technical and social challenges that Pratham Books faced. These need-finding interviews were crucial to understanding the key pain points while translators carry out their tasks and how a new interactive system can address those in an optimal manner.

- **Amateur Translators** – The Storyweaver interface of Pratham Books works on the shoulders of community translators who help with translating several of their books. While there are several in-house translators who are trained on the translation task, the majority of books get translated by community translators. These translators are mostly just altruistic bilingual speakers who are interested in translating the books in their mother tongue to provide children access to a wide volume of books in their mother tongues. While interacting with these translators, we found out that amateur translators may be translating in sub-optimal ways as compared to their professional counterparts. For example, they use openly available Machine Translation platforms such as Google or Bing Translate to get an MT output, paste it on the Storyweaver interface and then edit to get the final translation. While translating to a non-Latin script language, they also often rely on external apps for their transliteration needs. It can easily be seen how such a long-winded process can severely reduce the productivity of these community translators. Though this is mostly an issue of bringing multiple features in one interface, we gauge from our interviews that the choice of design can play a huge role in improving the process.
- **Creative Translations** – As opposed to the standard translation tasks where the primary aim is to accurately exchange information, the story writing/translating task requires creative and idiomatic use of sentences. MT systems are often trained on data that are factual/journalistic in nature and hence are largely incapable of generating creative translations on their own. Even when using MT in a post-editing setup, the somewhat frozen structure of the already output MT translations can inhibit translators from making it more creative. In such cases, it would be ideal to provide a broad range of diverse word-level suggestions which the translators can choose from.
- **Wide Array of Languages** – As already mentioned, Pratham Books publishes storybooks in a wide array of languages. There are some key issues with handling translations at the scale of 280 languages – (i) Most of these translations are carried out by bilingual speakers who might not be equally proficient in both languages. It is almost impossible logistically to obtain professional translators or even train new ones for the different language combinations. (ii) MT systems work satisfactorily only for a select few languages out of the box. The poor quality of translations beyond the high-resourced languages results in disfluent and incoherent sentence outputs. MT systems being trained in a similar manner to a language model are really good with next-word prediction. In this case, providing word-level suggestions can help translators make more appropriate choices.
- **Readability Grade** - When publishing children's storybooks, the level of reading can vary significantly depending on the

[3] https://storyweaver.org.in/



(a) The Encoder-Decoder setup which is modified such that instead of previous words, the input from the user is forced into the system.

(b) With every key push, the suggestions are shown to the users in two forms: full-sentence gist and two-word suggestions.

Figure 4: Free-text translation using Interactive NMT web application. When a character is entered by the user, it is sent as forced input to the Neural Machine Translation system (4a) and the output from that is displayed as gists and dropdowns (4b).

grade which the student is studying at. The level of reading, often known as Readability Grade, can differ in terms of the vocabulary (simple or complex), the concepts (easy or difficult), or the number of illustrations being used. Specific to our case, Pratham Books employs a 4-scale system[4] depending on the complexity of the text. Having such differentiation would mean that our proposed system should cater to the individual levels with different design choices and constrain the vocabulary of the system depending on the grade.

3.3 Previous Approaches

Previous approaches in MT have tried to address these challenges albeit independently. MT has often been criticized for being literal without accounting for the stylistic adaptations required for a particular task [65]. To address this, most of the previous approaches have worked towards style transfer for MT [75] such as controlling for poetic rhymes [34] or politeness [86] or developing personalized MT [81]. For controlling the readability level, previous works depend on constraining the vocabulary of translations [89] or have relied on post-hoc text simplification methods [100]. Factored MT [53] also has been a common approach to control translation outputs on rather fine-grained and granular factors. Such a method can, for instance, help with alleviating gender bias or fixing word sense disambiguation issues. One of the other factors while fixing for style is to account for the domain or the context of translations. Domain adaptation is a prevalent area within MT [51, 92], the primary goal of which is to introduce domain-specific translation vocabulary and has been mostly used for medical or legal document translations.

To achieve the end goal of "solving" translation, MT as a system should be able to carry out translations for several different combinations of languages. Performant MT systems require large amounts of data and hence it is not completely viable for language combinations where the available data is scarce. To counter this issue, often massively multilingual MT systems such as Aharoni et al. [3], Johnson et al. [44] are designed such that they are trained on multiple language combinations at once which in turn allow for learning shared multilingual representations. These shared representations generalize across multiple languages and hence compensating for the lack of data in low-resource languages [21, 26].

While these approaches have been researched purely from an MT perspective, they have not been tested in an interactive translation approach. Usually, the expectation in an interactive translation method is for the translator to provide their own style rather than enforcing it by controlling MT-generated translations.

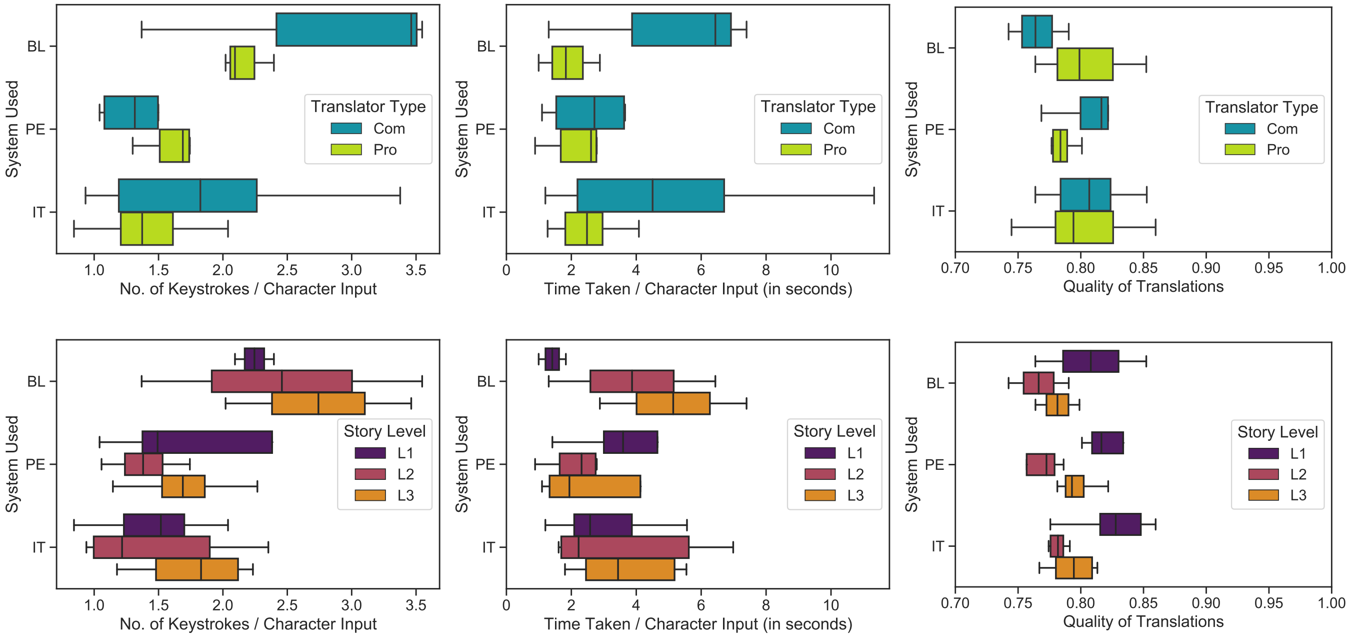
3.4 Our Approach

To tackle the aforementioned challenges, we would ideally want a system that allows for word-level suggestions which the translators could choose from if they would like to. We take the tried and tested route of guiding translators proactively through word-level suggestions. We use a prefix-based constrained decoding approach [49] and build an interface that provides a full-sentence gist as well as two-word suggestions in a dropdown format. We call this interface "INMT". Figure 4 shows how the interface looks.

Machine Translation is considered as a sequence-to-sequence task which traditionally uses recurrent neural networks (RNN) in an encoder-decoder combination. To enable an interactive setup, we take a constrained decoding setup wherein the user forces the MT to condition its current predicted word based on their partial input. Figure 4a shows how this setup works. For producing multiple suggestions based on partial inputs from the user, we rely on beam search decoding. It selects the most probable full translation for a given input sentence. If and when the translator diverges from this full translation, a new beam search is conducted from the partial input prefix till the end of the sentence is encountered.

As can be seen from the interface, we provide suggestions in two forms i.e. a full-sentence gisting and two-word dropdown suggestions. We alter the beam search parameters to achieve this. A full beam search (i.e. till the end of the sentence) is conducted only for the full sentence gisting. Whereas in the case of dropdown suggestions, we truncate the beam search at a length of 2 words. We preferred this design choice for dropdowns because decoding full-length translations through beam search lack diversity [35].

[4] https://storyweaver.org.in/reading_levels



(a) Number of Keystrokes (K) ↓ **(b) Time Taken (T) ↓** **(c) Quality of Translations (Q) ↑**

Figure 5: The performance of Interactive NMT on 3 different metrics – (a) Number of Keystrokes per character input (lesser is better) (b) Time Taken per character input (lesser is better) and (c) Quality of Translation (greater is better). The first row shows the difference across type of translators (Com (Community) or Pro (Professional) Translator) and second row shows the difference across story levels ordered by their difficulty (L1 < L2 < L3).

The pilot studies however revealed the actual benefit of such a design – (i) Providing a full-sentence gist meant that the translators could start right away without thinking/constructing a sentence in their head thus reducing the cognitive load (ii) A 2-word suggestion setup was ideal as too lengthy suggestions can be difficult to read in whole while too short of a suggestion would not give a sense of where the translation is going.

Through multiple design iterations and feedback from our pre-pilot study, we narrowed down to certain combinations of keystrokes which are naturally helpful for the translator while carrying out the translations.

- Tab : To get the next word from the selection.
- Enter ⟵ : To get all the words from the selection.
- ↑ ↓ : To alter the selection between multiple suggestions.
- Page ↑ Page ↓ : To traverse from one sentence to another.
- End : To end the translation process and download the translated document.

We also provide mouse click functionality. The user studies suggested that more experienced translators and computer users preferred to not take their hands off from the keyboard whereas the new translators tended to use only mouse clicks to navigate through constructing the translations.

We use OpenNMT [48], an open-source neural machine translation toolkit to build the MT system.[5] We write a new interactive translation mechanism to accept the user input and do constrained decoding, which is plugged on top of this toolkit. This helps in keeping up with the state-of-the-art models and other updates released through the toolkit and still keeping the interaction functional. OpenNMT being an established framework makes it easier for new models to be trained easily. If the target language uses a non-Latin script and users wish to use a Latin keyboard, the developer has to add an API to help with transliteration. The web interface employs REST API and sockets through which the real-time translations are carried out. With each keystroke from the user, a request is sent to the server and the existing request is invalidated. We tune the reaction time to around 200ms based on our pre-pilot study to allow users to adequately type their inputs before a request call is sent to the backend.

3.5 Feedback

We conduct multiple user studies to understand the effectiveness of mixed-initiative translations for this case. The user study is conducted in the same setup where the community translators contribute to translating storybooks for Pratham Books.

To get feedback on and improve our built system, we conducted multiple iterative pre-pilot, pilot, and main user studies. Here we describe our main representative user study which involved 10 translators working with Pratham Books and the feedback we received from them. We worked with only 10 translators as the overall process was rather long and required multiple interview sessions

[5] http://opennmt.net/

with them to gather qualitative feedback. While recruiting translators and designing our user study, we maximize the number of dimensions for which we could collect information. The dimensions considered in our study were as follows (I) 5 translators were professional and 5 were amateur (II) 3 different language combinations/directions (i) English to Hindi (ii) English to Tamil (iii) Hindi to English (of which we expand on (i)) and (III) 4 reading levels (reduced to 3). We compare the interactive translation suggestion approach with a baseline of providing no suggestion (BL) and with the Post-editing approach (PE) where a preliminary MT output is given such that the translators are required to edit and get the required translations out of it. To measure the efficacy, we record the performance of translators on 3 metrics viz. the number of keystrokes per character input/translation (K), the time taken for the same (T), and the quality of the overall translation (Q). We computed Q [range: 0 to 1] against the ground truth stories using LASER Embeddings [9] instead of the standard BLEU score metric in MT [78]. BLEU score is based on n-gram overlap of terms between the ground truth and projected translation, which does not function well in this case because the translations are extremely creative and therefore might have entirely different vocabulary although having the same meaning. The basic hypothesis is that K and T for mixed-initiative translations especially our interactive suggestions interface (IT) should be drastically lower than the baselines i.e. PE and BL. This is because suggestions are provided which makes it easier for the translators to select them and breeze through without typing anything. Figure 5 shows the performance of our system along the described 3 dimensions. In summary, we can see that Q improves a bit for IT when compared to BL and PE. Similarly, there is a decent reduction in K for IT in comparison to BL and PE. The T, on the other hand, does not alter much. We observed that this was most likely due to the translators' varying rates of learning effect with the interface. It was generally observed that the INMT/IT interface took more time to adapt as compared to PE. As the number of data points is small, we cannot draw any further conclusions as they would be statistically insignificant. Instead, we conducted 45-minute in-depth interviews with each translator to get their feedback on the different interfaces, their pain points, and what they would like to see improved. In the rest of this feedback section, we discuss both the qualitative and quantitative analysis of our study.

Overall the translators enjoyed the new interactive NMT interface for several different reasons. We discuss in further detail as follows:

- **Amateurs vs. Professionals** – We found that amateur translators have a very different requirement of suggestions than professionals. In the case of professional translators, they are already very adept at translating with very high accuracy and efficacy. They opined that suggestions on the side in form of bag-of-words would be much better than the current system of aggressively providing suggestions through dropdowns at every input, and hence preferred reactive suggestions i.e. provide suggestions only when needed. Whereas amateur translators enjoyed the proactive suggestions offered by the current system and wanted even more help through the system such as (an integrated dictionary) to help with their overall productivity. However, looking at the quantitative results, we can see that the professional translators were much more efficient (both K and Q) in using the IT system in comparison to the PE system. This can probably be attributed to their years of experience in translation and using such translation software. One other observation we made was that the professional translators were resistant to change from their already highly productive translation setup. For them adapting to a new interface would require a significant change in their workflow. We also saw a difference of behavior in the interaction input they used. Professional translators tended to use keyboards more as opposed to amateur translators who were using the mouse to select the suggestions and proceed with the translation as usual.
- **Help is not always required** – The suggestions provided by our system were useful to the translators but only some times. Through our interactions with the translators, we were able to understand how translators go about translating each sentence. The translator while thinking about a particular translation comes up with an ideal translation (which follows a certain sentence structure), before starting typing out on the computer. This occurred independently of the full sentence which was already getting suggested as a gist. If the suggestions agreed with this thought-out formation, it helped them with completing their thought process and the subsequent translation was carried out pretty quickly. However, if the suggestions countered their intuitive sense of translation, these suggestions completely threw them off and they had to think from scratch. Hence, there would be implications on when and how we show suggestions. The interfaces (especially the suggestions) in the future should be designed keeping this in mind. The suggestions should only nudge the translators when required not to force them.
- **Story Level** – The difficulty of conducting translations increases as the level of the story increases and hence suggestions should ideally be more appreciated as it is difficult and takes time to come up with a translation. This means that the interactive translation approach should ideally help the translators. However, contrary to our intuition, we can see that our approach was much more helpful to the translators for stories at initial levels. The reasons cited by the translators were that the higher-level stories required a creative edge which was missing from the suggestions. The suggested translations were literal and mostly on point.

3.6 Design Feedback

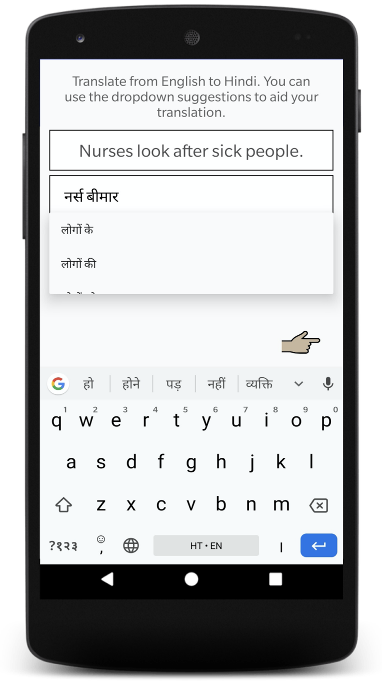
Other the preliminary feedback that we received regarding our interface, there were other observations which we made regarding how future interface can be designed.

- **Language of Translation** – The language pair between which the translation can severely influence the kind of interaction. One of the primary factors being the word order difference between two languages. Previous works have mostly worked around languages with a similar word order for eg. Translations between romance languages and English. For example, CASMACAT[4], a known tool in the Computer-Assisted Translation field, employs a phrasal matrix that is heavily reliant on having the same word order as that of the source sentence language. While probing for translators' preference for such a design, they pointed that such phrasal matrix will be confusing for language combinations where the word order is different such as in our case. It was also noted how having a different script can be problematic. We saw that X to English translations were much easier as compared to English to X where either there were keyboard issues or the transliteration which we used weren't good enough. In cases of transliteration suggestion being provided, should translation or transliteration be optimized for?

- **Sufficient Context** – Our initial prototypes used to display sentences to be translated individually. While this solves the job of translating, the translators often struggled with getting context in one go. We altered our interface to add the context of the story on the side such that the translators can read and re-read the story whenever required (when at a loss of context). However, there were further design considerations that were prompted by the translators. Instead of dividing the sentences into separate boxes, can the translations be done within the story paragraphs themselves? This could help with maintaining context at all times without requiring to shift the gaze often. In addition to this comment, translators suggested that the paragraph can be divided into a bunch of sentences (or phrases) rather than one sentence which can allow better context. These features are undergoing implementation and would require considerable thought processes such as how can a paragraph be divided in a way to maintain relevant context. Translators also pointed out that showing the individual illustrations can help with setting the context as is done in the Storyweaver interface. While showing illustrations is a specific case for our scenario, our interaction with translators revealed that in the existing systems, the text was always free-text and never accompanied by the visual context where the translation was being used.
- **Type of Device** – While conducting our user study, we noticed how the devices being used can significantly impact the experience of the translators. As the translator group is heterogeneous, we observed that they used laptops, desktops as well as IPads and mobile phones. The ideal types of interaction would change drastically between the types of devices. For example, desktop users were less likely to combine keyboard and mouse inputs when translating as compared to laptop users where the trackpad is more easily accessible. Similarly, keyboard inputs are tougher for mobile-like keyboards where touch-based inputs are more useful such as through intelligent/autocomplete suggestions that are shown on mobile keyboards.

4 II: CROWD-SOURCING FOR TRANSLATION DATA IN LOW RESOURCE LANGUAGES

The interaction with translators on our interactive translations approach revealed several areas where machine-aided translations can be helpful if designed carefully. In this case, we explore how a mixed-initiative setup can be beneficial in crowd-sourcing for translation data in low-resource languages. Machine-aided suggestions can move the focus from generating data from scratch to mere correction (post-editing) and annotation, which is easier and faster to accomplish.



Translate from English to Hindi. You can use the dropdown suggestions to aid your translation.

Nurses look after sick people.

नर्स बीमार

लोगों के

लोगों की

Figure 6: Mobile phone translations using the INMT-Lite Interface. It is designed to gather crowd-sourced translations in a quick way by providing suggestions at every step.

4.1 Background

Gondi is a language spoken by the 3-million strong Gond Community situated and spread across 8 states in Central India. However, compared to its counterpart languages spoken at a similar scale, Gondi is severely under-resourced in terms of data available. This can be attributed to several factors such as the language not being taught at school, the language not having its prevalence in print media as well as the non-standardization of the language itself due to its assimilation with native languages of different states. Although these causes are fairly responsible for the continued withering of the language, the ubiquity of language technologies and their applications have exacerbated this issue. Language technologies such as translation systems, social media apps, or even mobile keyboards mostly cater to only resource-rich languages which are thus forcing the youth of the community to migrate to these more popular languages. The most crucial requirement for building language technologies is the data. While data is not readily available in the case of Gondi, the community is very willing to crowd-source this data for the betterment of technologies especially translation systems which are important for information exchange. To enable faster and easier collection of data through crowd-sourcing, we want to understand how a mixed-initiative crowd-sourcing can work in this scenario.

4.2 Challenges

Our aim is to be able to collect data required for translation through the means of crowd-sourcing within the community. As crowd-sourcing is already a long and arduous process, we were interested in understanding how a mixed-initiative approach can work in this case. However, before developing any solution towards it, we worked along with the Gondi community to figure out the challenges which we would face when deploying an app for mixed-initiative crowd-sourcing.

- **Low-resource language** – As already stated, Gondi is a severely under-resourced language for one that is so widely spoken. Joshi et al. [46] in their analysis places Gondi as part of the class 0 set of languages which means that they are beyond redemption unless there is a focused effort put in reviving the resources for that language. While there have been previous works that have aimed at collecting data in Gondi through crowdsourcing, they are pretty slow in obtaining translations which means that it will take a long time to bridge this gap [67]. Without enough data to train the MT systems, the generated translations can be subpar to the extent that some of them can even be incomprehensible. Here is where taking a mixed-initiative approach can help. The translations produced by suboptimal MT systems are usually good to the next word though are incapable of producing full-length coherent translations. These word-level predictions can thus be used to provide suggestions to the users.
- **Lack of Personal Computers** – Situated in the tribal regions of India, the Gond community is off the radar from having access to technological advances such as personal computers. This also means that the community is mostly digitally illiterate. However, the Indian telecom market has suddenly boomed with the introduction of cheap mobile internet and phones. This has led to mobile phones proliferating to every nook and corner of India and becoming a ubiquitous device among households. This means that mobile phones can be used for the purpose of effective crowd-sourcing of translation data. Being handheld and thus available at the disposal whenever required makes it convenient for users to use it at any time. Moreover, there is a huge potential of gamification of the process such as using in a language learning setup such as in the case of Duolingo [95] and Google Bolo.
- **Network Connectivity** – Even though mobile phones are ubiquitous among these communities and are served by a decent network for voice calls, the network bandwidth, and overall connectivity is of poor quality. This issue can be a bottleneck for dispensing real-time suggestions through network calls which is essential for the current mixed-initiative approaches to translation. One of the ways to mitigate this issue is to build offline models which can allow for a model to be deployed into the user's phone.

4.3 Previous Approaches

Crowd-sourcing for language resource development has been a well-studied area of research and work [71, 104]. Much of it has depended on offering monetary incentives via platforms such as Mechanical Turk [16]. However, there are significant limits to the data acquired in this manner, such as no incentives to collect data for low-resource languages [25] or to collect data that is globally representative due to scaling issues. To address such issues, attempts have been made to collect data through alternative incentives such as through citizen science [1, 29, 67]. Mixed-Initiative approaches have also been used recently to collect data where annotators are primed by an AI agent to quicken the annotation process [7, 70, 87]. This approach has been extended to active learning setups where the collected data is used to train the model on the fly [6, 79]. To counter limited storage space, easy installation across multiple devices as well as for real-time ML applications, ML models are sometimes deployed through the browsers [88]. Such a mechanism is also useful to counter network connectivity issues. However, larger models are required to enable better predictions and hence these local models are often used in a federated learning approach where there is a regular update of the global model based on the gradients from the local model [101].

4.4 Our Approach and Feedback

We extended the INMT platform by taking into account the key challenges which this scenario posed. INMT-lite is an android application and framework developed to train and develop a mobile version of the original INMT system described in §3. INMT-lite was designed with two unique features: (i) Ability to use the INMT tool with a low-resource device like a smartphone. This included designing a user interface for smart-phone that allowed for the collection of translation data in a gamification setup. (ii) Allow usage of the tool in a low network bandwidth environment by deploying an offline MT model on the phone. The initial design considerations included dropdown suggestions similar to previous approaches (as shown in Figure 6). The most challenging issue was of deploying a fully capable MT model to the smartphone. As OpenNMT at the time only included state-of-the-art models which are heavy on memory and latency, we took a basic approach of constructing a sequence-to-sequence Recurrent Neural Network (RNN) from scratch using tf-lite [88]. This model was trained with 20,000 Hindi-Gondi sentences and with 360,000 English-Hindi sentences to compensate for the extremely low-resource Hindi-Gondi combination. The model size was approximately 70MB as opposed to the original 1GB models and hence could easily fit in a mobile device. While gathering feedback from the community, it was understood that displaying suggestions in a bag-of-word manner was more desirable to enable more seamless interaction between typing and selecting suggestions. Their suggestion on the design choice was understandable as it was mentioned that the suggestions were used in a passive sense as the suggestions sometimes were incorrect and not the words which they desired for.

5 III: INCORPORATING VISUAL CONTEXT DURING TRANSLATION

We learned from our usual interactions with translators that providing as much background as possible may lead to better translations. Visual context is the most easily captured of all the contexts and is rarely employed in current translation settings. During our general interactions with translators, we realized that providing as much context as possible can lead to better translations. Of all the contexts, visual context is the most easily capturable one which is scarcely used in the current translation setups. Through this case, we explore how mixed-initiative translations can be helpful to conduct translations taking visual context into account.

5.1 Background

Rekha[6] leads a Language Service Provider (LSP). LSP is an entity that offers services related to languages and this mostly includes translation and localization of content to multiple languages. The whole process of translation includes multiple stages starting from listing down the client's expectations, preparing a quote, then preparing the documents before the translation takes place, handing the documents over to multiple translators in their network, collecting the translations back, and getting it proofread and finally delivered. Rekha is frequently collecting feedback to understand how the process can be smoothed out for translators and clients. There are a couple of issues that exist with the current setup which we discuss as part of the challenges.

5.2 Challenges

One of the key pain points which they face while translating is the lack of visual context while translating the documents. Often the translators are tasked with translating free-flowing text devoid of the document structure. This is further amplified for cases like webpage localization where the texts that are required to be translated are less than 5 words such as "About Us", "Events" which can have different meanings in varied contexts. Moreover, contextual information such as the length of the translation required, or understanding pictorial references are not available and can lead to suboptimal or erroneous translations. This is where translating the document in-situ can help. Not only does this help with error-free translation, but also allows eliminates the copy-editing segment of the translation process which can be often tedious. We outline and describe these challenges in more detail as follows:

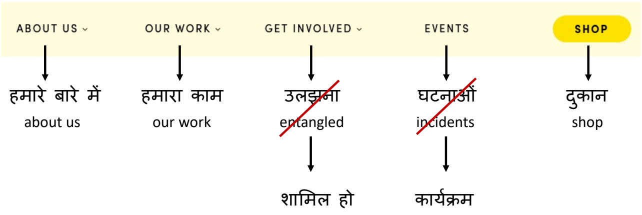


Figure 7: Words or phrases taken out of context can lead to erroneous translations. As seen here, the translation API does not where and how the word is being used (no visual context) and hence can provide with any one of the possible number of synonymous translations of a particular word.

- **Lack of Visual Context** – It is sometimes difficult to understand the context in which a particular translation is required. It is even harder to provide these abstract contexts to an MT system. Figure 7 shows an example where sentences taken out of context can result in erroneous translation. Given that the words appear on a menu bar, they have certain meanings associated with it. As can be seen, the translation API produces translations without understanding the context of where it is being used. This results in erroneous translations which do not fit the given context, but can be fixed by a human mediator who can select one of the many synonymous translations produced.

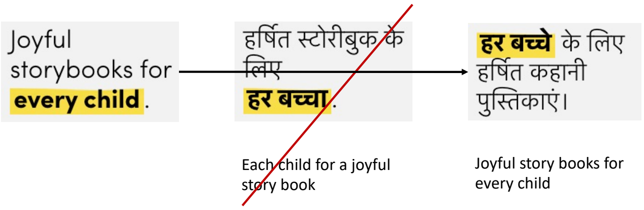

Joyful storybooks for every child.
~~हर्षित स्टोरीबुक के लिए हर बच्चा.~~
~~Each child for a joyful story book~~
हर बच्चे के लिए हर्षित कहानी पुस्तिकाएं।
Joyful story books for every child

Figure 8: Not taking the structural context (in this case, the phrase highlight) into account produce incorrect translations. In such cases, human intervention is required to fix the formatting in the correct way to show the translated phrase in the most optimal way possible.

- **Preserving Webpage Structure** – HTML documents are inherently structured documents which means that the elements of the page are divided into blocks. Such a block system is used for purposes like formatting the document consistently. The existing translation systems mostly work take free-text as input and are not able to parse such structured formats easily. When these structured phrases are converted to free text form to carry out translations separately, the HTML structure is lost which can sometimes result in erroneous translations. Figure 8 shows an example of the same where translating between languages with different word orders can be difficult. Translation is carried out for the following text in form of a nested HTML element "Joyful storybooks for every child" where there is an formatting/styling emphasis on "every child". With automatic translation, each element gets translated individually which results in a flawed sentence structure which when back-translated means "Each child for a joyful storybook".

5.3 Previous Approaches

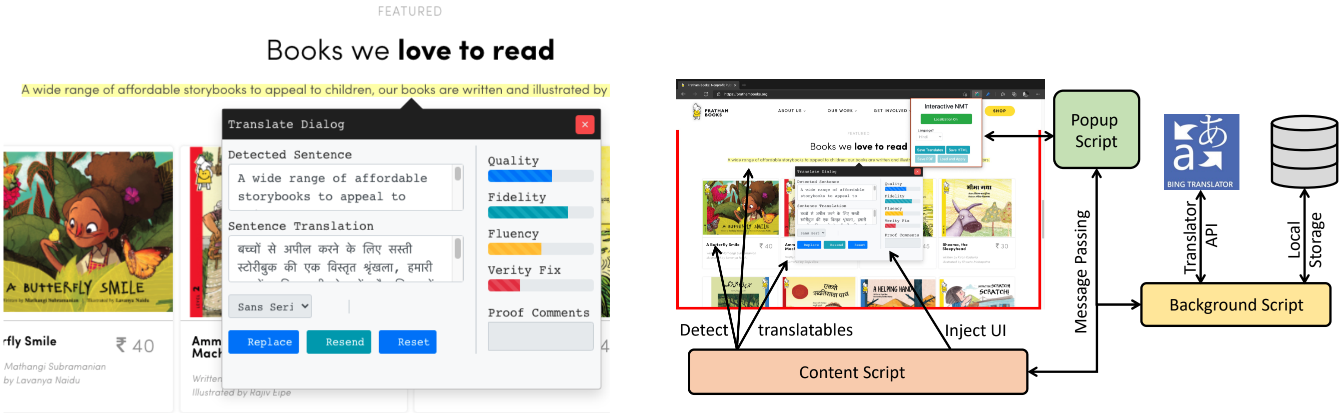
Previous approaches to localization have often involved large-scale translation through online MT systems. However, in addition to the lack of visual context, there is almost no appropriate sentence context as well. Most of the sentences/phrases for such cases are very short [72]. Hence, most of the efforts have included humans in the loop at some point such as through post-editing [31] or through suggestions [64].

Providing visual context has also been tackled before. Popular machine translation software such as SDL Trados, Screen Match [55] rely on displaying screenshots for the corresponding translations. Leiva and Alabau [62] shows how in-situ translations can help immensely with accurate contextual translations [82]. Lack of such visual context can lead to several translation issues like misgendering (English is non-gendered, other languages are), misinterpretation of words if appearing more than once and other such issues [72] [90].

5.4 Our Approach and Feedback

We build an interface through the means of a browser extension (javascript) that allows for such in-situ translations of web pages. Figure 9a shows the interface carrying out an example translation. The translator will be able to select any text element present on the webpage including menus and buttons which will then popup a modal. The modal includes the already translated output based

[6] name fictionalized



FEATURED

Books we **love to read**

A wide range of affordable storybooks to appeal to children, our books are written and illustrated by

Translate Dialog

Detected Sentence

A wide range of affordable storybooks to appeal to

Sentence Translation

बच्चों से अपील करने के लिए सस्ती स्टोरीबुक की एक विस्तृत श्रृंखला, हमारी

(a) The Interactive NMT Browser Interface. Users can select specific text on the webpage which needs to be translated.

(b) Working of the Interactive NMT Browser Plugin. This plugin works completely on browser as an extension making the installation easy.

Figure 9: Webpage translation using Interactive NMT Browser Plugin. The visual context provided by the surrounding texts and images enables translators to interactively fix the automatic free-text translations obtained from a translation API.

on the present sentence which the translator can choose to change. In addition, the modal provides multiple metrics to judge the final translations. The extension can work on any webpage. Once a translator is finished with the translations, they will be able to send it to the LSP in form of a markup which can further be proofread/analyzed, and finalized. In addition to helping with visual context, there are some unique benefits that such a system offers. For example, there is no longer a need for the internationalization of web pages thus making them universally applicable. Also, given that this application runs entirely within the browser without reliance on any external specialized software, there is huge potential to democratize translation collection.

The initial qualitative feedback which we received through a pilot interview for our system was favorable. We showed the app to a different LSP from the one with whom we did the need-finding interview. They noted how prior approaches to localization were time-consuming since they were required to capture screenshots for each instance that requires translation. Often, there were brief phrases like "Read More," which required more visual context, such as placement on the page than just screenshots or words. They suggested that such a system could also be ported for PDF document translations where such visual contexts can be helpful for plot legends or translating annotated figures.

6 RECOMMENDATION FOR MT AS A SOCIO-TECHNICAL SYSTEM

6.1 Mapping Socio-Demographic Contexts with Technical Needs

It is important to identify different socio-demographic contexts where technology use can differ. Through our paper, we do the same by highlighting three use-cases that make use of Machine Translation but in 3 entirely distinct ways altogether. To re-iterate, we saw how MT is used as a suggestion tool (i) to boost the productivity of amateur translators, (ii) to collect translation data for low-resource languages, and (iii) to aid with highly contextual translations during webpage localization. It is already clear how diverse, though, extremely specific MT applications can be. The use-cases that we cover are fairly limited as compared to the many challenges which exist in the real world. There are several other communities which would require MT in a unique setup to address these challenges [91]. For instance, as already discussed in our introduction, how can we facilitate a smoother use of MT in contexts such as during crisis which handles just bilingual speakers who have probably never translated, who are often remote and the translation process demands for a faster turnaround. There are many more conceivable socio-demographic situations or obstacles while adopting MT, and it would be nearly impossible to list them all in one go. However, a plausible solution to this problem is to have a coordinated effort between the NLP and HCI + CSCW + DEV communities to design and develop human-centered MT. It is especially crucial for the NLP community to be cognizant of the societal needs of translation and not be a victim of the McNamara Fallacy i.e. relying solely on metrics in complex situations and losing sight of the bigger picture. It is important to understand that the translation and exchange of information is the end-goal of the process and everything else such as boosting quality for MT does not mean anything if they do not work in the real world.

6.2 Optimizing Machine Translation for different metrics

In NLP and other similar data-driven fields like Machine Learning and Computer Vision, there are often benchmarks that allow for comparison between the performances of different models. While such goal-setting has resulted in enormous advancement in the field, such a race (and flag-planting) has can be harmful in a variety of ways. First, as Ethyarajah and Jurafsky (2020) [28] suggest, there is a utility mismatch between the leaderboard and the practitioners in the field. For instance, an NLP practitioner can care about the latency or the robustness of the model while the leaderboards focus mostly on ranking through accuracy and quality while not capturing the metrics which can be useful in the real world. Choudhury and Deshpande discuss how leaderboards tend to reward the models which perform on the bases of sheer size and often disregard how it can prove to be unfair for a subset population where it is being deployed to [19]. When it comes to diversity and practical usefulness of the metrics, Machine Translation (MT) fares much better compared to other NLP tasks. The tasks released by WMT (Workshop for Machine Translation) are often used to judge the quality of MT models, which include several metrics of evaluation. Furthermore, every iteration of WMT introduces new language pairs in their shared tasks including low-resourced languages, such as English-Kazakh, English-Lithuanian, English-Gujarati in WMT 2019. There are also several other tasks where the MT models are evaluated such as the robustness, automatic post-editing, etc. It is known that automatic metrics such as BLEU and METEOR are often bad proxies of actual quality. Thus, even when a system performs as good as human translators on these metrics (that is to say "the system achieves human parity"), actual human evaluation of the system often shows that they are much inferior to human translations [17, 59]. This is sufficiently realized by the community and there is a quality metric hunt every year which looks for new metrics that correlate with human performance better.

As already discussed through our work, we observed that the majority of the state-of-the-art models were not suitable for our scenario. To obtain relatively acceptable translation quality, these models required huge quantities of data (millions of parallel sentences per a language pair) and were computational and memory intensive. Developing the optimal MT model was an iterative and time-consuming process of experimenting with several models and finally narrowing down to one. Aside from size and latency, there are numerous other metrics that might have real-world societal implications, such as the uncertainty and interpretability of the model prediction, bias in the model, or robustness[83]. Re-iterating the example which we took in the earlier section of the paper, in a high stakes situation such as conversing with law enforcement it is desirable to understand the uncertainty of the predicted translation as opposed to a benign task like ordering food. Fortunately, leaderboards in NLP and related areas are working towards integrating specialized metrics such as model efficiency [69], robustness [24, 54] and social biases [73, 74]. Similar to our observations, it was observed for EfficientQA that the model architecture of 5MB models differed significantly in comparison to the ones which were unrestricted in size.

6.3 Better Mixed-Initiative Translations

Previous approaches in mixed-initiative translations including ours have mostly looked at how machines can help humans or vice-versa (aka machine- or human-in-the-loop) while achieving the common goal of translation. However, one of the lesser-explored directions in mixed-initiative translation interfaces is that of how both humans and machines can learn during the process thereby setting up an alternate personal incentive. Interactive Machine Learning and Active Learning are areas that look at machine learning through human interaction. Specific to NLP, there have been attempts at a machine that learns language through games which humans play [97, 102] as well as active learning approaches in several NLP tasks such as MT [36, 40]. In a social scenario, such a setup will allow the model to learn over time without re-training the model for new instances. Some such examples include active learning through crowd-sourcing for low-resource data [6] or mitigating social biases in the model through crowdsourcing [105]. In our experience of deploying models for Gondi, we saw that collection of data can help underpowered MT systems (trained only on 20,000 sentences).

Humans can also learn while interacting with machines. For instance, Duolingo collects translation data while users are learning a new language through the app [95]. Using assistive writing tools can also enable humans to learn language passively through interaction with suggestions. These suggestions can help with discovering new idioms and commonly used phrases in a particular language [20]. From a social perspective, we observed through our field studies that the younger generation of the Gondi Community is not fully aware of certain lost words in Gondi which were replaced with borrowed words from other languages. Having a machine-in-the-loop system can help with getting hold of new words in the Gondi language which can also help with language rejuvenation. Future work can also focus on building seamless interfaces for collecting multilingual data through video games or systems like reCAPTCHA [96]. It can be otherwise difficult to collect such data at scale even with crowd-sourcing mechanisms. Such data is important to bridge the resource gap in NLP without which there is a risk of many languages getting extinct [46].

7 CONCLUSION

In this paper, we propose how language translation should be treated as a socio-technical system as there is a sufficient overlap between translation, people, and the community at large. One of the ways to achieve this is to make use of existing approaches like mixed-initiative interactions which already work at the interface between people and technology. We describe through multiple socially motivated use-cases how the mixed-initiative translations can not only help with solving some of the technical challenges but also aid with mitigating some of the social challenges. While we limit ourselves to the three use-cases, we believe that they are sufficiently broad to cover several subsets of communities that face similar social challenges. Through our observations and interactions with multiple communities, we also put forward some recommendations which can further help with bridging the gap between the challenges faced by all four pillars of a socio-technical system.

ACKNOWLEDGMENTS

This work would not have been possible without the contributions of many people who helped with conducting user research, designing, constructing, and deploying the systems, and offering input at various phases of the project. We would like to thank Christain Barnes in particular for his assistance in conducting the pilot user research with the Pratham Books translators. We also thank Anamika Radhakrishnan, Purvi Shah, Amna Singh, Sumana M, and several of Pratham Books' in-house and community translators for the initial inspiration of the work and for assisting us all along the way, from providing key insights into how their translation process works to helping us in gathering resources for conducting user studies and interviews. We thank Devansh Mehta, Shubranshu Choudhary, and others at CGNet Swara and the Gond community who have been influential in taking the direction of using our system in a crowd-sourcing set up to collect data for a low-resource language. We thank Rakhi Sundar and Santosh Kale from Microsoft for providing several insights into LSPs and designing translation interfaces. We thank Arul Menezes, Niranjan Nayak, and Sundar Srinivasan from Microsoft Research and Microsoft Bing for their feedback on improving the translation interfaces. We also thank Mohd Sanad Zaki Rizvi for their key contributions to the INMT-lite model.

REFERENCES

[1] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. 2020. Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2819–2826. https://www.aclweb.org/anthology/2020.lrec-1.343

[2] Mark S. Ackerman. 2000. The Intellectual Challenge of CSCW: The Gap between Social Requirements and Technical Feasibility. *Hum.-Comput. Interact.* 15, 2 (Sept. 2000), 179–203. https://doi.org/10.1207/S15327051HCI1523_5

[3] Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3874–3884. https://doi.org/10.18653/v1/N19-1388

[4] Vicent Alabau, Christian Buck, Michael Carl, Francisco Casacuberta, Mercedes García-Martínez, Ulrich Germann, Jesús González-Rubio, Robin Hill, Philipp Koehn, Luis Leiva, Bartolomé Mesa-Lao, Daniel Ortiz-Martínez, Herve Saint-Amand, Germán Sanchis Trilles, and Chara Tsoukala. 2014. CASMACAT: A Computer-assisted Translation Workbench. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, 25–28. https://doi.org/10.3115/v1/E14-2007

[5] Jeffrey Allen. 2003. Post-editing. *Benjamins Translation Library* 35 (2003), 297–318.

[6] Vamshi Ambati. 2012. *Active Learning and Crowdsourcing for Machine Translation in Low Resource Scenarios*. Ph.D. Dissertation. Carnegie Mellon University, USA. Advisor(s) Carbonell, Jaime. AAI3528171.

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. *Guidelines for Human-AI Interaction*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[8] Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the Translation Initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.nlpcovid19-2.5

[9] Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics* 7 (March 2019), 597–610. https://doi.org/10.1162/tacl_a_00288

[10] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

[11] Lois Bastide. 2018. *Crisis communication during the Ebola outbreak in West Africa: The paradoxes of decontextualized contextualization*. Springer International Publishing, Cham. https://archive-ouverte.unige.ch/unige:107132 ID: unige:107132.

[12] Ludwig von Bertalanffy. 1968. General systems theory as integrating factor in contemporary science. *Akten des XIV. Internationalen Kongresses für Philosophie* 2 (1968), 335–340.

[13] Erik Brynjolfsson, Xiang Hui, and Meng Liu. 2019. Does machine translation affect international trade? Evidence from a large digital platform. *Management Science* 65, 12 (2019), 5449–5460.

[14] Jody Byrne. 2007. Caveat translator: Understanding the legal consequences of errors in professional translation. *Journal of Specialised Translation* 7 (2007), 2–24.

[15] Patrick Cadwell. 2015. *Translation and trust: a case study of how translation was experienced by foreign nationals resident in Japan for the 2011 great east Japan earthquake*. Ph.D. Dissertation. Dublin City University.

[16] Chris Callison-Burch. 2009. Fast, Cheap, and Creative: Evaluating Translation Quality Using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 286–295. https://aclanthology.org/D09-1030

[17] Arun Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 643–653. https://doi.org/10.18653/v1/P18-1060

[18] Albert Cherns. 1976. The Principles of Sociotechnical Design. *Human Relations* 29, 8 (1976), 783–792. https://doi.org/10.1177/001872677602900806 arXiv:https://doi.org/10.1177/001872677602900806

[19] Monojit Choudhury and Amit Deshpande. 2021. How Linguistically Fair Are Multilingual Pre-Trained Language Models? *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 14 (May 2021), 12710–12718. https://ojs.aaai.org/index.php/AAAI/article/view/17505

[20] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[21] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747

[22] Alan Cooper. 2004. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2nd Edition)*. Pearson Higher Education.

[23] Sven Coppers, Jan Van den Bergh, Kris Luyten, Karin Coninx, Iulianna van der Lek-Ciudin, Tom Vanallemeersch, and Vincent Vandeghinste. 2018. Intellingo: An Intelligible Translation Environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174098

[24] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2020. RobustBench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670* (2020).

[25] Sandipan Dandapat and William Lewis. 2018. Training deployable general domain mt for a low resource language pair: English–bangla. (2018).

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[27] José Esteban, José Lorenzo, Antonio S. Valderrábanos, and Guy Lapalme. 2004. TransType2 - An Innovative Computer-Assisted Translation System. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. Association for Computational Linguistics, Barcelona, Spain, 94–97. https://www.aclweb.org/anthology/P04-3001

[28] Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the Eye of the User: A Critique of NLP Leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4846–4853. https://doi.org/10.18653/v1/2020.emnlp-main.393

[29] James Fiumara, Christopher Cieri, Mark Liberman, and Chris Callison-Burch (Eds.). 2020. *Proceedings of the LREC 2020 Workshop on "Citizen Linguistics in Language Resource Development"*. European Language Resources Association, Marseille, France. https://aclanthology.org/2020.cllrd-1.0

[30] Michael Fleming and Robin Cohen. 2000. Mixed-initiative translation of Web pages. In *Proceedings of the Fourth Conference of the Association for Machine Translation in the Americas: Technical Papers*. Springer, Cuernavaca, Mexico, 25–29. https://link.springer.com/chapter/10.1007/3-540-39965-8_3

[31] Raymond Flournoy and Christine Duran. 2009. Machine translation and document localization at Adobe: From pilot to production. *MT Summit XII: proceedings of the twelfth Machine Translation Summit* (2009), 425–428.

[32] Raymond S Flournoy and Chris Callison-Burch. 2001. Secondary benefits of feedback and user interaction in machine translation tools. In *Workshop paper*

for "MT2010: Towards a Roadmap for MT" of the MT, Summit, Vol. 8. Citeseer, 2–3.

[33] Lena Foljanty. 2015. Legal Transfers as Processes of Cultural Translation: On the Consequences of a Metaphor. *Kritische Vierteljahresschrift für Gesetzgebung und Rechtswissenschaft* 2 (2015), 89–107.

[34] Dmitriy Genzel, Jakob Uszkoreit, and Franz Och. 2010. "Poetic" Statistical Machine Translation: Rhyme and Meter. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA, 158–166. https://www.aclweb.org/anthology/D10-1016

[35] Kevin Gimpel, Dhruv Batra, Chris Dyer, and Gregory Shakhnarovich. 2013. A Systematic Exploration of Diversity in Machine Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1100–1111. https://www.aclweb.org/anthology/D13-1111

[36] Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Avignon, France, 245–254. https://www.aclweb.org/anthology/E12-1025

[37] Spence Green, Jason Chuang, Jeffrey Heer, and Christopher D. Manning. 2014. Predictive Translation Memory: A Mixed-Initiative System for Human Language Translation. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology* (Honolulu, Hawaii, USA) *(UIST '14)*. Association for Computing Machinery, New York, NY, USA, 177–187. https://doi.org/10.1145/2642918.2647408

[38] Spence Green, Jeffrey Heer, and Christopher D. Manning. 2013. The Efficacy of Human Post-Editing for Language Translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 439–448. https://doi.org/10.1145/2470654.2470718

[39] Ana Guerberof Arenas. 2009. Productivity and quality in MT post-editing. (2009).

[40] Gholamreza Haffari, Maxim Roy, and Anoop Sarkar. 2009. Active Learning for Statistical Phrase-based Machine Translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Boulder, Colorado, 415–423. https://www.aclweb.org/anthology/N09-1047

[41] Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, et al. 2018. Achieving human parity on automatic chinese to english news translation. *arXiv preprint arXiv:1803.05567* (2018).

[42] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[43] W. John Hutchins. 2004. The Georgetown-IBM Experiment Demonstrated in January 1954. In *Machine Translation: From Real Users to Research*, Robert E. Frederking and Kathryn B. Taylor (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 102–114.

[44] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics* 5 (2017), 339–351. https://doi.org/10.1162/tacl_a_00065

[45] Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung Challenges of Building and Deploying Language Technologies for Low Resource Language Communities. In *Proceedings of the 16th International Conference on Natural Language Processing*. NLP Association of India, International Institute of Information Technology, Hyderabad, India, 211–219. https://www.aclweb.org/anthology/2019.icon-1.25

[46] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[47] Martin Kay. 1997. The Proper Place of Men and Machines in Language Translation. *Machine Translation* 12, 1/2 (1997), 3–23. http://www.jstor.org/stable/40009025

[48] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of ACL 2017, System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 67–72. https://www.aclweb.org/anthology/P17-4012

[49] Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Proceedings of the Association for Machine Translation in the Americas*. 107–120.

[50] Rebecca Knowles, Marina Sanchez-Torron, and Philipp Koehn. 2019. A user study of neural interactive translation prediction. *Machine Translation* 33, 1-2 (2019), 135–154.

[51] Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain Control for Neural Machine Translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*. INCOMA Ltd., Varna, Bulgaria, 372–378. https://doi.org/10.26615/978-954-452-049-6_049

[52] Philipp Koehn. 2009. A Web-Based Interactive Computer Aided Translation Tool. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*. Association for Computational Linguistics, Suntec, Singapore, 17–20. https://www.aclweb.org/anthology/P09-4005

[53] Philipp Koehn and Hieu Hoang. 2007. Factored Translation Models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, Prague, Czech Republic, 868–876. https://www.aclweb.org/anthology/D07-1091

[54] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. 2020. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv preprint arXiv:2012.07421* (2020).

[55] Geza Kovacs. 2012. ScreenMatch: providing context to software translators by displaying screenshots. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. 1375–1380.

[56] Stacey Kuznetsov. 2006. Motivations of Contributors to Wikipedia. *SIGCAS Comput. Soc.* 36, 2 (June 2006), 1–es. https://doi.org/10.1145/1215942.1215943

[57] Philippe Langlais, George Foster, and Guy Lapalme. 2000. TransType: a Computer-Aided Translation Typing System. In *ANLP-NAACL 2000 Workshop: Embedded Machine Translation Systems*. https://www.aclweb.org/anthology/W00-0507

[58] Samuel Läubli, Chantal Amrhein, Patrick Düggelin, Beatriz Gonzalez, Alena Zwahlen, and Martin Volk. 2019. Post-editing Productivity with Neural Machine Translation: An Empirical Assessment of Speed and Quality in the Banking and Finance Domain. In *Proceedings of Machine Translation Summit XVII Volume 1: Research Track*. European Association for Machine Translation, Dublin, Ireland, 267–272. https://www.aclweb.org/anthology/W19-6626

[59] Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 4791–4796. https://doi.org/10.18653/v1/D18-1512

[60] Anne-Marie Laurian. 1984. Machine Translation : What Type of Post-Editing on What Type of Documents for What Type of Users. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stanford, California, USA, 236–238. https://doi.org/10.3115/980491.980542

[61] H J Leavitt and B M Bass. 1964. Organizational Psychology. *Annual Review of Psychology* 15, 1 (1964), 371–398. https://doi.org/10.1146/annurev.ps.15.020164.002103 arXiv:https://doi.org/10.1146/annurev.ps.15.020164.002103

[62] Luis A. Leiva and Vicent Alabau. 2014. The Impact of Visual Contextualization on UI Localization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 3739–3742. https://doi.org/10.1145/2556288.2556982

[63] William Lewis. 2010. Haitian Creole: How to Build and Ship an MT Engine from Scratch in 4 days, 17 hours, & 30 minutes. In *Proceedings of the 14th Annual conference of the European Association for Machine Translation*. European Association for Machine Translation, Saint Raphaël, France. https://www.aclweb.org/anthology/2010.eamt-1.37

[64] Donghui Lin, Yoshiaki Murakami, Toru Ishida, Yohei Murakami, and Masahiro Tanaka. 2010. Composing Human and Machine Translation Services: Language Grid for Improving Localization Processes. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA), Valletta, Malta. http://www.lrec-conf.org/proceedings/lrec2010/pdf/317_Paper.pdf

[65] Adam Lopez and Matt Post. 2013. Beyond bitext: Five open problems in machine translation. In *Twenty Years of Bitext*.

[66] Elliott Macklovitch. 2006. TransType2 : The Last Word. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. European Language Resources Association (ELRA), Genoa, Italy. http://www.lrec-conf.org/proceedings/lrec2006/pdf/14_pdf.pdf

[67] Devansh Mehta, Sebastin Santy, Ramaravind Kommiya Mothilal, Brij Mohan Lal Srivastava, Alok Sharma, Anurag Shukla, Vishnu Prasad, Venkanna U, Amit Sharma, and Kalika Bali. 2020. Learnings from Technological Interventions in a Low Resource Language: A Case-Study on Gondi. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources

Association, Marseille, France, 2832–2838. https://www.aclweb.org/anthology/2020.lrec-1.345

[68] Alan K. Melby, Melvin R. Smith, and Jill Peterson. 1980. ITS: Interactive Translation System. In *COLING 1980 Volume 1: The 8th International Conference on Computational Linguistics*. https://www.aclweb.org/anthology/C80-1064

[69] Sewon Min, Jordan Boyd-Graber, Chris Alberti, Danqi Chen, Eunsol Choi, Michael Collins, Kelvin Guu, Hannaneh Hajishirzi, Kenton Lee, Jennimaria Palomaki, et al. 2021. NeurIPS 2020 EfficientQA Competition: Systems, Analyses and Lessons Learned. *arXiv preprint arXiv:2101.00133* (2021).

[70] Robert R. Morris, Mira Dontcheva, and Elizabeth M. Gerber. 2012. Priming for Better Performance in Microtask Crowdsourcing Environments. *IEEE Internet Computing* 16, 5 (2012), 13–19. https://doi.org/10.1109/MIC.2012.68

[71] Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, Los Angeles, 122–130. https://aclanthology.org/W10-0719

[72] Victor Muntés-Mulero, Patricia Paladini Adell, Cristina España-Bonet, and Lluís Màrquez. 2012. Context-Aware Machine Translation for Software Localization. In *Proceedings of the 16th Annual conference of the European Association for Machine Translation*. European Association for Machine Translation, Trento, Italy, 77–80. https://www.aclweb.org/anthology/2012.eamt-1.15

[73] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[74] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 1953–1967. https://doi.org/10.18653/v1/2020.emnlp-main.154

[75] Xing Niu, Marianna Martindale, and Marine Carpuat. 2017. A Study of Style in Machine Translation: Controlling the Formality of Machine Translation Output. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2814–2819. https://doi.org/10.18653/v1/D17-1299

[76] Sharon O'Brien and Federico Marco Federici. 2019. Crisis translation: Considering language needs in multilingual disaster settings. *Disaster Prevention and Management: An International Journal* (2019).

[77] Minako O'Hagan. 2011. Community Translation: Translation as a social activity and its possible consequences in the advent of Web 2.0 and beyond. *Linguistica Antverpiensia* 10, 2011 (2011), 11–23.

[78] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[79] Álvaro Peris and Francisco Casacuberta. 2018. Active Learning for Interactive Neural Machine Translation of Data Streams. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Brussels, Belgium, 151–160. https://doi.org/10.18653/v1/K18-1015

[80] Jaana Porra and Rudy Hirschheim. 2007. A lifetime of theory and action on the ethical use of computers: A dialogue with Enid Mumford. *Journal of the Association for Information Systems* 8, 9 (2007), 29.

[81] Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized Machine Translation: Preserving Original Author Traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 1074–1084. https://www.aclweb.org/anthology/E17-1101

[82] Antonio Rizzo and Marco Palmonari. 1998. Context and Consciousness: Activity Theory and Human Computer Interaction, Bonnie A. Nardi (ed.). *User Modeling and User-Adapted Interaction* 8, 1-2 (1998), 153–157.

[83] Sebastin Santy and Prasanta Bhattacharya. 2021. A Discussion on Building Practical NLP Leaderboards: The Case of Machine Translation. *arXiv preprint arXiv:2106.06292* (2021).

[84] Sebastin Santy, Sandipan Dandapat, Monojit Choudhury, and Kalika Bali. 2019. INMT: Interactive Neural Machine Translation Prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*. Association for Computational Linguistics, Hong Kong, China, 103–108. https://doi.org/10.18653/v1/D19-3018

[85] Rico Sennrich. 2018. Why the time is ripe for discourse in machine translation. In *Second Workshop on Neural Machine Translation and Generation*.

[86] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Controlling Politeness in Neural Machine Translation via Side Constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 35–40. https://doi.org/10.18653/v1/N16-1005

[87] Sanket Shah, Pratik Joshi, Sebastin Santy, and Sunayana Sitaram. 2019. CoSSAT: Code-Switched Speech Annotation Tool. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*. Association for Computational Linguistics, Hong Kong, 48–52. https://doi.org/10.18653/v1/D19-5907

[88] Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, Stan Bileschi, Michael Terry, Charles Nicholson, Sandeep N. Gupta, Sarah Sirajuddin, D. Sculley, Rajat Monga, Greg Corrado, Fernanda B. Viegas, and Martin Wattenberg. 2019. TensorFlow.js: Machine Learning for the Web and Beyond. Palo Alto, CA, USA. https://arxiv.org/abs/1901.05350

[89] Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical Machine Translation with Readability Constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Linköping University Electronic Press, Sweden, Oslo, Norway, 375–386. https://www.aclweb.org/anthology/W13-5634

[90] Huatong Sun. 2001. Building a Culturally-Competent Corporate Web Site: An Exploratory Study of Cultural Markers in Multilingual Web Design. In *Proceedings of the 19th Annual International Conference on Computer Documentation* (Sante Fe, New Mexico, USA) *(SIGDOC '01)*. Association for Computing Machinery, New York, NY, USA, 95–102. https://doi.org/10.1145/501516.501536

[91] Mustapha Taibi and Uldis Ozolins. 2016. *Community translation*. Bloomsbury Publishing.

[92] Zhaopeng Tu, Yang Liu, Zhengdong Lu, Xiaohua Liu, and Hang Li. 2017. Context Gates for Neural Machine Translation. *Transactions of the Association for Computational Linguistics* 5 (2017), 87–99. https://doi.org/10.1162/tacl_a_00048

[93] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[94] Lucas Nunes Vieira, Minako O'Hagan, and Carol O'Sullivan. 2020. Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases. *Information, Communication & Society* 0, 0 (2020), 1–18. https://doi.org/10.1080/1369118X.2020.1776370 arXiv:https://doi.org/10.1080/1369118X.2020.1776370

[95] Luis von Ahn. 2013. Duolingo: Learn a Language for Free While Helping to Translate the Web. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces* (Santa Monica, California, USA) *(IUI '13)*. Association for Computing Machinery, New York, NY, USA, 1–2. https://doi.org/10.1145/2449396.2449398

[96] Luis von Ahn, Benjamin Maurer, Colin McMillen, David Abraham, and Manuel Blum. 2008. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science* 321, 5895 (2008), 1465–1468. https://doi.org/10.1126/science.1160379 arXiv:http://www.sciencemag.org/content/321/5895/1465.full.pdf

[97] Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning Language Games through Interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 2368–2378. https://doi.org/10.18653/v1/P16-1224

[98] Brian Whitworth, Jerry Fjermestad, and Edward Mahinda. 2006. The Web of System Performance. *Commun. ACM* 49, 5 (May 2006), 92–99. https://doi.org/10.1145/1125944.1125947

[99] Joern Wuebker, Spence Green, John DeNero, Saša Hasan, and Minh-Thang Luong. 2016. Models and Inference for Prefix-Constrained Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 66–75. https://doi.org/10.18653/v1/P16-1007

[100] Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing Statistical Machine Translation for Text Simplification. *Transactions of the Association for Computational Linguistics* 4 (2016), 401–415. https://doi.org/10.1162/tacl_a_00107

[101] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. 2019. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 13, 3 (2019), 1–207.

[102] Xingdi Yuan, Marc-Alexandre Côté, Jie Fu, Zhouhan Lin, Chris Pal, Yoshua Bengio, and Adam Trischler. 2019. Interactive Language Learning by Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2796–2813. https://doi.org/10.18653/v1/D19-1280

[103] NA Yurko, IM Styfanyshyn, UM Protsenko, and Yu R Slodynytska. 2020. Tourism translation: the key peculisrities. (2020).

[104] Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing Translation: Professional Quality from Non-Professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 1220–1229. https://aclanthology.org/P11-1122

[105] Honglei Zhuang and Joel Young. 2015. Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (Shanghai, China) *(WSDM '15)*. Association for Computing Machinery, New York, NY, USA, 243–252. https://doi.org/10.1145/2684822.2685301

[106] Álvaro Peris, Miguel Domingo, and Francisco Casacuberta. 2017. Interactive neural machine translation. *Computer Speech & Language* 45 (2017), 201–220. https://doi.org/10.1016/j.csl.2016.12.003