Web Scraping Localized Parallel Multilingual Help Content in Indian Languages

S. Winston Cruz
Department of Linguistics, KIKS
University of Mysore
Mysuru, India

G. Roch Libia Rani
Department of Computer Applications
De Paul College
Mysuru, India

ABSTRACT
The need for multilingual corpora has witnessed a quantum leap with the development in web mining and large language models (LLM). Multilingual data extraction from websites is a way of developing parallel corpora. Controlled use of web scraping is a useful technique for the creation of this corpus. Among the various types of localized content on the web, including the machine translations, content produced engaging human translators and reviewers are most useful followed by machine translated content that has undergone human post editing. The help center documents and the terms and conditions documents available on the websites in different languages come under these categories. In this paper, such content is manually identified and the issues in scraping them are discussed. A Python code that uses BeautifulSoup library for extracting these materials in various Indian languages like Hindi, Kannada and Tamil is presented. The concerns related to arranging the content parallelly with their source in English is then discussed. Finally, details of the sample parallel corpus extracted is analyzed and presented.

General Terms
Parallel corpora, Indian languages, web mining

Keywords
Web scraping, BeautifulSoup, localization, Tamil, Kannada, Hindi

1. INTRODUCTION
Web scraping is the extraction of data from websites, portals etc. automatically and classifying and presenting the data again according to a desired template defined by chosen fields and purpose. It is done by using bots and other programs that extract the data automatically to store them in the target files or locations.

Web scraping multilingual content has not been very popular in India. Of late, training AI and LLM models for translation has increased the need for multilingual corpus [1]. Different agencies have started scraping comparable content in various languages. They are used in developing parallel corpus to train their AI models in both processing language content as well as translating between languages.

Web scraping as an idea may sound as a simple ‘copy and paste’ exercise of any web content [2]. However, technical nuances built into it differentiates it from a normal action of copying and pasting any digital content. It differs mainly because of the intent for which it is done. It also varies because of the technology it is engineered with.

2. INTENT AND TECHNOLOGY
When material on the web is scraped, there is an intent behind it. For instance, if terms and conditions of user agreements are collected through web scraping, it may be for comparing the terms and conditions and studying the various clauses for their inclination towards benefitting either of the parties involved in signing the agreement.

If the scraped content is the various product descriptions from different online retail portals, it may be for comparing the price or features of the various products across these portals. It may become a tedious experience for users looking forward to a simple review of the product to view the product pages on the original retail portals. A properly scraped and precisely presented content could be more useful to specific users than the original pages.

Technically, scraping includes web crawling too. It is also called data extraction or web mining. It also uses parsing. There could be analysis and summarization of the content involving other tools too. Nowadays, these are mostly done by AI tools, especially those that are built from data provided by LLMs. In return, web scraping, as a method, helps in the building of these LLMs.

When large amounts of data are scraped from different websites and portals, they could be used for machine learning (ML) purposes. Of late, technical advancements suggest that even smaller sized data can become helpful if appropriate technologies of deep learning (DL) are used [3]. Sometimes, scraped content is also used to generate synthetic data for use by the LLMs. When multilingual content is stored and if they are comparable content, they may be used for studying the translation patterns, quality, accessibility etc. Such content could also be used for studying the structure of the content and reviewing their usability, reader friendly aspects etc. Therefore, web scraping can also be called data mining for a specific purpose. Presentation of the data in a user-friendly way is facilitated here.

Copying and pasting a content would mean a manual effort. In web-scraping computer programs that use software libraries to do it directly from websites are used. Therefore, web scraping is mostly automatic along the guidelines in a controlled and targeted manner. They are not random scraping in any case unless otherwise that is the intention. In some cases, the crawlers and scrapers are programmed to restrict themselves to or have fixed target websites.

3. ISSUES
3.1 Technical issues
Leaving aside the important discussions on ethical issues of web scraping, the technical issues themselves can be plenty depending upon the structure of the source websites, the nature of the content, the tools or programs used etc. For instance, many articles on a newspaper site may need paid subscriptions. Only authorized users who have logged in officially will be able to view the entire content. Some websites differentiate

between genuine users and bots and block the latter [4]. They may use technology like OTP or captcha etc. for the purpose. Simple web scraping of the content may not be possible in these cases.

Initially, web scraping started with extraction of information or content from static pages which did not pose much issues. Nowadays, web pages have a lot of dynamic information compiled according to various triggers including tracked user activities. Content is also presented through animations, in response to mouse hovering actions, or as truncated and expandable pages etc. They are not simple HTML pages. For instance, the section where the terms and conditions of the National Highways Authority of India website is presented [5], which is available in both Hindi and English, may be using angular or react type of client-side rendering through JavaScript. BeautifulSoup cannot be directly used to scrap content here. It is best suited for static HTML. The websites similar to NHAI's may need different scraping techniques and programming for extracting content like the use of libraries like Selenium or Playwright.

Sometimes, websites may not follow any particular template to arrange their multilingual content. These could make the extraction of the content difficult across the different language pages through preset templates of scraping. Even when such content is scraped, there could be issues in classifying them and storing it according to the new presentation template. The scraped content may become unusable for any further processing.

However, in essence, web scraping is about creating a structure and a template for presenting unstructured data that are available across pages. Therefore, while the variety of formats in which content is available across websites may prove to be challenging, they do not discourage programmers from scraping them. Time and resources are spent to extract these contents and present them or analyze them according to the desired format which is actually the purpose of scraping. It is in this sense that web scraping is also viewed as web harvesting. Thus, the attempt may add both a new purpose and some clarity to the cluttered content.

3.2 Script and language issues in scraping Indian language content

One of the technical issues in scraping multilingual Indian language content is the use of the same script by more than one language. This makes language identification the first issue in scraping contents, especially for processing linguistic material [6]. If one takes the major Indian languages, even among the officially recognized languages, many like Hindi, Marathi, Bodo, Dogri etc. use the same Devanagari character set. Some languages which are at different stages of recognition but have a recognizable presence on the web also use character sets of other languages. For instance, Rajasthani uses the Devanagari script. Some languages, like Tulu, a Dravidian language, use the scripts of their respective state language. Tulu, spoken in the state of Karnataka in India, is written using the script of Kannada, the official language of the state. Content is found in more than one script for languages like Sindhi, Manipuri etc. Of late, the Government of Manipur has adopted Mayek as the official script for Manipuri (Meitei). Students learn it in school and it is being promoted extensively by the state in the public space. However, historically, Bengali script was used till recently in print media for Manipuri. All these, and also the mixing of codes in some cases like the social media texting, makes language identification a core issue in multilingual web scraping. This leads to extra effort in identifying the languages where the scripts are similar or the same [7].

3.3 To include or not to include machine translated content?

Extracting multilingual data usually includes machine translated content too. However, machine translated and synthetic content are not the preferred data for use in machine learning (ML) systems. Especially, the performance of LLMs trained on synthetic data in downstream tasks like machine translation or summarization may be unreliable [8].

In some cases, it may not be possible to conclude if the parallel content in Indian languages is human translated or machine translated or machine translated and post edited (MTPE) by human translators. Some websites offer a mixture of both [9]. On the one hand, what makes it indecisive could be viewed as a proof for the improved efficiency of machine translation systems which produce human-like translations in terms of readability, of late. The issues of translation in such pages could also be because of ineffective post-editing. There are many examples of both kinds available on the web as far as Indian languages are concerned. In this study, we have left out obviously identifiable machine translated content.

3.4 The websites and apps offering multilingual content in Indian languages

There are many websites that offer multilingual content in Indian languages. Many service providers offer the option of changing the language of the interface of their websites and mobile apps in different Indian languages. The content is either pre-translated and presented or translated by tools on the fly and presented. Machine translators that translate on the fly are mostly AI powered these days. Google Cloud Translation API, which offers translation between nearly 200 languages, can be incorporated into any website through subscription [10]. Microsoft Translator API powered by their Azure AI services [11], Amazon Translate, powered by Amazon Web Services [12] etc. can all provide real-time translation of website content giving an advantage of reaching to offshore clients and users.

Many mobile phone companies provide different language interfaces for their phones. These include many Android based phones as well as the iOS based products of Apple. The phones offer accessibility in different languages chosen according to the locale where they are sold. They also have input methods for the different languages including many Indian languages. The interface content available in Indian languages could be a good source of verified translations. A few mobile applications provide interfaces in Indian languages [13]. Some of these include apps from social media companies to online grocery retailers. Some of them, even if they do not have different language interfaces, at least allow searching in Indian languages. The number of languages seems restricted to less than 10 including Indian English in many cases. This study has restricted itself only to multilingual content available on websites and not mobile OS or app interfaces.

Some websites declare that they offer their content in multiple languages. However, they update their content in the main language page but may not update the other language pages immediately [14, 15, 16, 17]. There could be different reasons for the phased updating like availability of monetary resources or localization resources etc.

4. NEED FOR LANGUAGE INDEXING

There seems to be no stated policy available about the inclusion of languages in the various products of the software and tech

companies. Microsoft has been offering Indian language interfaces for its Windows OS for a long time. They have advanced features of reviewing in Indian languages with dictionary support in their products like their MS-Office packages. The Windows OS language packs which include fonts and keyboard interfaces are available for more than 20 Indian languages. The Tamil language pack is available for four locales which include India, Malaysia, Singapore and Sri Lanka. In the Tamil (India) pack, Microsoft has included the Tamil-99 input keyboard developed by the Government of Tamil Nadu also [18]. This feature is not offered by most other software providers who cater to a large variety of world languages. However, Microsoft's homepage [19] allows interface change only to four Indian languages apart from the 'India (English)' option. The website does allow modifying the feed to Tamil language. However, it has to be found by searching in the browser or by other means rather than accessing the Tamil page from the homepage by changing the language option. This scenario may sound strange to a common user as Microsoft Bing Translator, a popular MT by Microsoft offers Tamil machine translation options. In fact, Bing Translator is the first one to offer a unique option of fetching translations in three different 'tones' of standard, casual and formal Tamil [20]. Some parts of the Microsoft website do offer parallel content in different languages like the 'Microsoft Services Agreement' page. It is available in 11 Indian languages, including Hindi, Kannada and Tamil [21].

Google Search, the popular search engine, on the other hand, provides the page in many Indian languages. The language can be selected directly just below the search bar on the home page [22]. Google also has absolutely parallel localized service terms and conditions and other help center articles in different Indian languages which can be selected through a drop-down menu on the footer of the page [23]. Like the 'Microsoft Services Agreement' [21], Google pages are also parallel texts to their respective English pages both in terms of the content as well the structure. Each line of the Indian languages pages here are directly comparable to their English source. Texts formatted like this facilitate developing parallel corpus in these languages using simple methods [24].

The unavailability of policy statements on language interfaces or new language inclusion for different interfaces or pages in many websites create hurdles for web scrapers of multilingual content to arrive at a consistent template. Occasionally there are press releases or new items [13]. However, such information need verification.

5. METHODOLOGY

The target of the present study is to check the possibilities of extracting parallel localized multilingual help guides and terms and conditions documents available on different websites in different Indian languages and analyze the trends in localization available through the extracted content.

This research specifically concentrated on extracting Hindi, Kannada and Tamil content that could be arranged parallelly against their English source. The first task is to produce a list of websites which contain the content that is required in any one source language. This list should then be expanded to include webpages with comparable content with that of the source pages. Extracting this list may be done by executing a smaller program written either in Python or other similar languages. All these lists of web pages should then be subjected to a filtering out process where the back references links and other general links should be removed. Writing a code to extract the content of the web pages from the lists is the third task. Writing the extracted content to a desired template and saving it for further processing is the next task. This process too could be contained in the same program that was written to extract the content. The content could possibly be saved as a database or in rows and columns in a simple MS-Excel file. The decision will depend on the NLP or other tools that will be used to analyze the extracted comparable content. Arranging the content from multiple languages comparably is the fifth task. The arrangement must facilitate the analysis of the content especially with NLP tools. The analysis of the content could be done in different ways depending upon the aim and scope of the research.

In the present research, the websites which offered multilingual help center and terms and conditions content were manually identified. A preliminary code that would extract all the html links from an Indian language home page of the help center or a subsection of it was written in Python and used. The links were manually scrutinized to remove the non-help center related and the general links available on the page. The Tamil, Kannada and Hindi pages links were finalized in that order first and only those links which had their pair in each language were retained and others were edited out. This task too was done by manual scrutiny of the content. In the next step, these links were compared with the English page links and were vetted in the same manner. Thus, the links for all the four languages from which content will be extracted were finalized. To extract content from the vetted list of links, a Python code that uses the BeautifulSoup library was generated through relevant prompts in the ChatGPT code generator engine [25]. The program was run in the Google Colab environment [26] and links in each language file were accessed by the program and the content were scraped. The extracted content files get added to the cloud space. These were downloaded and stored in the desired location. Various methods, including the use of built in excel formulae, could be used to combine data of different languages into one sheet.

6. THE STRUCTURE OF CODE

The structure of the code is like the following:

```
The preliminaries
Declaration of input (with the list of
  multilingual user guide URLs to be
  scraped) and output files
Reading URLs from text file with line strip
Function to extract paragraphs from user
  guide pages using BeautifulSoup
  Locating and leaving out other links in
  non-main content region
  Extracting main content paragraphs
Creating or opening Excel workbook,
  appending extracted data and saving
```

7. DISCUSSION AND ANALYSIS

7.1 General Factors

The complexity of extracting various Indian language data differs between websites. Since, the present study is a limited one restricted only to the help center documents and terms and conditions of multilingual websites, script related language identification issues discussed in a previous section was overcome by restricting ourselves to URLs that contain a language parameter where the language is declared. Shifting between different language versions was easy in some websites, for example, Apple, Google and YouTube. We had to only change the language code in the language parameter holder. For example, extracting English content from the

YouTube URL requires only 'hl=en' to be changed to 'hl=as', 'hl=hi' or 'hl=ta' for extracting the same page in Assamese, Hindi or Tamil languages respectively [27].

However, language parameter change is blocked by some websites [9]. Such sites may set cookies for the session with the particular language chosen while logging in or visiting the page first. Any change in the language parameter in the URL does not affect the display language. Changing the language on the URL, for example, from 'hi' (Hindi) to 'kn' (Kannada), does not change the display as the session language chosen in the beginning may override the new request. This can either be overcome by resetting the cookies or changing the language manually using the language option on the page.

As mentioned before, some websites had parallel content in Indian languages which were not comparable to their English source as only the latter had been updated recently. The English page [14] and its equivalent Hindi, Kannada and Tamil [15, 16, 17] pages are an example for this. The English page has been updated in September 2025 whereas the Hindi, Kannada and Tamil pages were last updated in February 2023. A few other websites, including that of the online retailer Flipkart discussed here, had Indian language pages which had Indian language file names in the respective script in their URLs for each language. These also had to be manually listed for the program for scraping their content.

7.1.1 Coding Factors

The coding had to be simple for using the BeautifulSoup library. The program extracted the paragraphs and listed them against the file names. Initially, some issues were noticed with linked texts appearing without space and text after some punctuations like a colon (:) appearing without space. These were corrected after a test run. A time.sleep of 0.25 seconds delay between requests was maintained to avoid any jamming.

7.2 The iPad User Guide Corpora

A total of 3,640 pages or files of the iPad User Guide [28] were scraped during this study with 910 files each in English, Hindi, Kannada and Tamil. The details are given in Table 1.

The parallel corpora created by web scraping here were not subjected to any detailed corpus cleaning processes. The words themselves have roughly been defined as the textual material occurring between spaces. Number words have also been included in the number of total words. For instance, zip codes are counted as words in this corpus. If the zip code is written with any spaces, it would be counted as two words. Trademarks and product names were given in English itself in the Indian language pages. These were not deleted and were scraped and saved under the respective language corpus. No other substantial normalization processes were applied to the corpora. Yet, the approximate statistics reveal some general trends in the localization industry as well as some universal properties of the languages represented in the corpora.

7.2.1 Localization Trends

In Table 1, we can see that the number of scraped segments of the three Indian languages is almost the same. This could mean that they have been created or updated about the same time. The English corpus has about 1000 segments more the Indian languages. A cursory look at English corpus shows that there could have been few updates in it and some added steps [29] here and there are identified in certain processes which are missing in the Indian language corpora. For instance, content extracted from one of the English pages [30] shows the following three steps whereas its corresponding Indian language page content [31, 32, 33] have only the first:

Ex. 1. a) Add a text box: Tap.
b) Add text inside a shape: Tap.
c) Add a sticky note: Tap. – English [30]

a) ___,या पर टैप करें। - Hindi [31]

a) ___, ಅಥವಾಅನ್ನು ಟ್ಯಾಪ್ ಮಾಡಿ. – Kannada [32]

a) ___, அல்லது -ஐத் தட்டவும். – Tamil [33]

Such instances in the corpus could be either because of phased localization by the content provider or because of retrieval issues in the execution of the extraction code.

In some places, two segments showing two steps in the English corpus have been combined into one in the Indian language pages. For example, the two steps, 'Go to the Settings app on your iPad.' and 'Tap Privacy & Security.' from English get combined as 'Go to the Settings app on your iPad, tap Privacy & Security.' in the Indian language corpora. There are also segments in the Indian language pages which are unavailable in the English pages. This could mean that these steps or descriptions were found redundant and were edited out in the English guide pages or the product itself has undergone some update and the processes have been trimmed which have not been updated in the Indian language guide pages. While iPad User Guide pages have almost negligible amounts of such instances, it is quite common to see such trends in any localization projects related to tech products.

7.2.2 Linguistics Factors

Since the research is about language corpus, many interesting linguistic factors emerge if one were to study the corpus in a detailed manner. As mentioned previously, the extracted multilingual corpus also has scope for translation factors to be analyzed. These factors can studied both for their prominence and function in their individual languages as well as in comparison with other languages.

7.2.2.1 Postpositions and Agglutination

Among the languages, Kannada and Tamil are agglutinative languages [34], that is, many grammatical features, case markers, auxiliary verbs etc. appear in the form of suffixes attached to the stem of the words unlike in English and Hindi where, for example, the pre/postpositions appear as words separated by a space. The same postpositions in Kannada and Tamil are written together with the noun. Generally, a certain morpho-phonemic changes the words have unique forms and are listed as different tokens by simple programs for splitting words. For example, in the scraped corpora of the iPad User Guide pages, there are common, everyday Kannada syntactic constructions like *gurutisuvavarege* and *vyatyaasaveenu* used in the following pages:

Table 1. English iPad User Guide - segments and words

Languages	Paragraph segments (approx.)	Number of words (approx.)	Unique words (approx.)
English	19,950	291,500	6,820
Hindi	18,930	361,700	6,640
Kannada	18,950	261,600	16,510
Tamil	18,950	233,500	20,580

Table 2. English iPad User Guide Top 25 Words

Sl. No.	Word Token	Freq.	Sl. No.	Word Token	Freq.
1.	the	17366	14.	of	3729
2.	to	12509	15.	app	3179
3.	tap	9061	16.	with	2333
4.	you	8120	17.	go	2273
5.	your	7225	18.	for	2232
6.	a	6664	19.	Apple	2224
7.	and	6032	20.	see	1917
8.	on	5978	21.	if	1771
9.	or	5683	22.	an	1723
10.	then	4686	23.	use	1690
11.	in	4627	24.	screen	1579
12.	iPad	4353	25.	when	1513
13.	can	3843		**Total**	**122310**

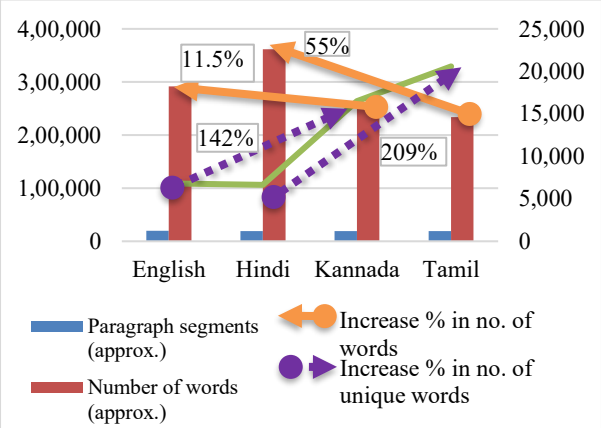

Figure1. Increase directions and percentages

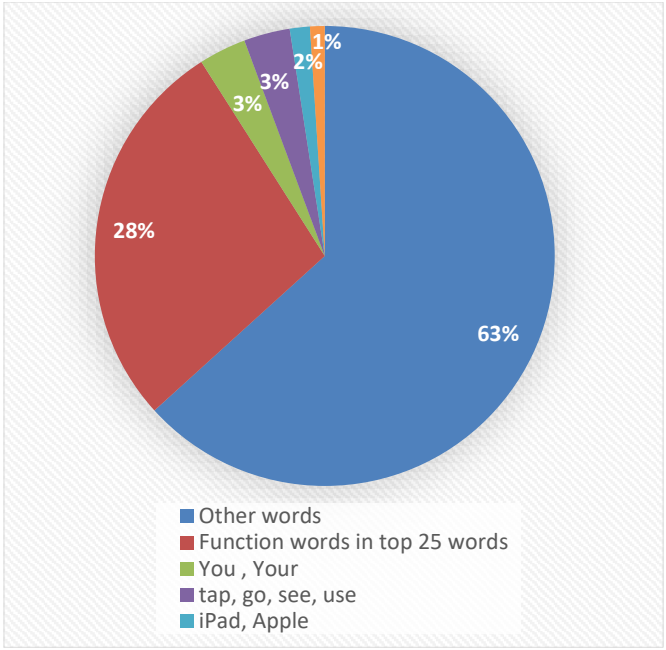

Figure2. English iPad User Guide - top 25 words statistics

Ex. 2. iPad won't dim or lock as long as it detects attention.' – English [35]

तो iPad तब तक डिम या लॉक नहीं होगा जब तक आप इसकी ओर देखते हैं। - Hindi [36]
to iPad tab tak Dim yaa laak nahiin hoogaa jab tak aap iski aur deekhtee hain.
ಗಮನವನ್ನು ಗುರುತಿಸಿರುವವರೆಗೆ iPad ಡಿಮ್ ಅಥವಾ ಲಾಕ್ ಆಗುವುದಿಲ್ಲ. – Kannada [37]
gavanavannu gurutisuvavarege iPad dim adhava laak aaguvadilla

Ex. 3. What is the difference between iMessage, RCS, and SMS/MMS? [38]
iMessage, RCS ಮತ್ತು SMS/MMS/RCS ನಡುವಿನ ವ್ಯತ್ಯಾಸವೇನು? [39]
iMessage, RCS mattu SMS/MMS/RCS naDuvina vyatyaasaveenu

The equivalents English expressions in the corpus for *gurutisuvavarege* and *vyatyaasaveenu* are 'as long as it detects' and 'what is the difference' respectively. Each of these 'single' words in Kannada are translations of 5 and 4 English words respectively. At the sentence level, the 11-word English clause in ex. 2 is translated into Hindi, the other non-agglutinative language in the corpora, as a 17-word clause. The same has been translated into Kannada with 7 words.

7.2.2.2 Statistics of Segments and Words

In the present study, overall, we can notice that the number of words of the non-agglutinative languages like English and Hindi show almost 11.5% (between Kannada and English) to 55% increase (Tamil and Hindi) in the iPad User Guide pages. However, the agglutinative languages show a reversal trend when it comes to the unique words in the corpora. The number of unique words of the agglutinative languages show an increase of 142% (between English and Kannada) to 209% (Hindi and Tamil). Such statistics allow the project managers to keep a tab on consistency of the linguistic trends in the localization works.

7.2.2.3 Function Words and Content Words

Many other linguistic trends become visible even with broad analysis of the corpus. The top 25 frequent words account for around 40% of the corpus of 291,500+ word English corpus extracted here. Like any other English corpus function words and pronouns are more frequent in this present study also. More than 15 of the top 25 words are functors or pronouns. The pronouns 'you' and 'your' alone occur some 15,345 times. Typical to the description and instructions about digital products and contents, four verbs, namely 'tap, go, see, use' appear 14,941 times. As the extracted content is Apple User Guide, two proper nouns 'iPad' and 'Apple' are found over 6500 times. Two common nouns, 'app' and 'screen' also occur about 4750 times.

Table 3. Tamil iPad User Guide Top 15 Words

Sl. No.	Word Token	Freq.
1.	உங்கள் *ungal*	5276
2.	தட்டவும் *taTTavum*	5050
3.	என்பதைத் *enpatait*	4747
4.	அல்லது *alladu*	4685
5.	நீங்கள் *niingaL*	4636
6.	தட்டி *taTTi*	3587
7.	உள்ள *uLLa*	3441
8.	-ஐத் *-ait*	2343
9.	மற்றும் *maRRum*	2276
10.	Apple	2163
11.	iPadஇல் *ipaDil*	2153
12.	ஒரு *oru*	1873
13.	செய்யவும் *ceyyavum*	1829
14.	செல்லவும் *cellavum*	1786
15.	iPad	1570

The Indian language corpora (Hindi, Kannada and Tamil) also reveal similar interesting figures and analysis. In the Tamil iPad User Guide, for instance, *ungal* 'your' in the genitive form tops the list of frequent words followed by *taTTavum* 'tap', the verb in the imperative form. With different case markers or verbal declensions, the total instances of these words increase to more than double their numbers. One can notice from the table that *niingal* 'you' in the nominative form has more than 4000 occurrences. There are more than 25 forms of the word *ungal* 'your' in the Tamil corpus and together with the nominative, they appear more than 10,000 times in the corpus. It can also be noticed that *taTTavum* 'tap – hortative' and *taTTi* 'tap – adverbial' appear in the top 15 words list. Together, they account for more than 8000 occurrences. In fact, there are more than 50 forms of the word *taTTu* 'tap – imperative' are found in the user guide. All these suggest that a separate detailed study is needed to reflect on their various aspects of the Indian language corpora and NLP tools that are specifically designed for individual Indian languages are need to do a thorough analysis of the corpora.

8. CONCLUSION

Extracting multilingual content which is parallel in nature is possible through relatively simple measures as shown here. Especially, as the content targeted here was restricted to that of the help guide in some multilingual websites of tech companies. Scraping remained straightforward through some manual filtering for structure and presentation. Since the extracted content was either human translated or reviewed, it is more useful for training and testing MT systems, especially those that are used in localization projects. The parallel corpora offered a preview of the statistical trends in localizing from English to Indian languages. Key differences were noticed in statistics between a non-agglutinating language like Hindi and agglutinating languages like Kannada and Tamil. The agglutinating languages had a smaller number of total words but more unique words or word forms. If more statistical insights from localization projects are produced from a wider localized content, some benchmark statistical scales could be arrived at which will help the overall quality of localization projects in Indian languages at a pan-India level. Such content also provides a base for a deeper analysis of the nature of the localized content itself like style, readability, use of terminology, transliteration techniques, common errors, ambiguity etc. A more detailed analysis of the language content can reveal techniques of handling various case-endings of nouns, paradigms of verb inflections, splitting or combining sentences etc. in Natural Language Processing (NLP).

A thorough analysis of the scrapped content can also lead to improvements in scrapping of Indian language content in terms of language identification, alignment with the source, parsing, chunking etc. The alpha-syllabic nature of the Indian writing system, their Unicode arrangements could be analyzed in comparison with world language character sets and more sophisticated and dynamic scrapping techniques can be devised for multilingual content in general and Indian languages content in particular. Such studies in the future can even improve the very presentation of Indian languages on the web and add economy and efficiency to the processes.

9. REFERENCES

[1] Shaharbanu, A., & McDonald, S. (2025, 08 01). *Legality of data scraping under Indian law*. India Business Law Journal. Retrieved 10 30, 2025, from https://law.asia/india-data-scraping-regulation/

[2] Lotfi, C., Srinivasan, S., Ertz, M., & Latrous, I. (2022). Web Scraping Techniques and Applications: A Literature Review. In R. Pal & P. K. Shukla (Eds.), *SCRS Conference Proceedings on Intelligent Systems* (pp. 381-394). Soft Computing Research Society. https://doi.org/10.524 58/978-93-91842-08-6-38

[3] Gupta, P., & Jamwal, S. S. (2025). Enhancing NLP for Low-Resource Language by Developing Deep Learning-Powered Morphological Analysis of Dogri: An End-to-End Pipeline from Corpus Construction and Linguistic Annotation to Model Training and Deployment. *SN Computer Science*, *6*. https://link.springer.com/article/10.1007/s42979-025-04429-9

[4] Bale, A. S., Ghorpade, N., S, R., Kamalesh, S., R, R., & S, R. B. (2022). Web Scraping Approaches and their Performance on Modern Websites. In *2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC)* (pp. 956-959). IEEE. 10.1109/ICESC54411.2022.9885689

[5] NHAI, Ministry of Road Transport and Highways. (n.d.). *Terms & Conditions*. National Highways Authority of

India. Retrieved November 18, 2025, from https://nhai.gov.in/#/terms-conditions

[6] Agarwal, M., Alam, M. M. I., & Anastasopoulos, A. (2023). LIMIT: Language Identification, Misidentification, and Translation using Hierarchical Models in 350+ Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14496–14519). Association for Computational Linguistics.

[7] Ingle, Y., & Mishra, P. (2025). ILID: native script language identification for Indian languages. *arXiv*, *2507.11832v2*. arXiv. https://doi.org/10.48550/arXiv.2507.11832

[8] Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z., & Tao, H. (2024). Synthetic data in AI: Challenges, applications, and ethical implications. *arXiv pre-print arXiv*, *2401.01629*.

[9] Amazon.com, Inc. (n.d.). *Online Shopping site in India: Shop Online for Mobiles, Books, Watches, Shoes and More*. Amazon.in. Retrieved November 18, 2025, from http://www.amazon.in

[10] Google. (n.d.). *Cloud Translation*. Google Cloud. Retrieved September 15, 2025, from https://cloud.google.com/translate

[11] Microsoft. (n.d.). *Translator Text API*. Microsoft. Retrieved September 17, 2025, from https://www.microsoft. com/en-us/translator/business/translator-api/

[12] Amazon Web Services. (2025). *Amazon Translate API Reference - Amazon Translate API Reference*. AWS Documentation. Retrieved November 18, 2025, from https://docs.aws.amazon.com/translate/latest/APIReference/welcome.html

[13] Times of India. (2024, June 24). Flipkart launches support for Tamil, Telugu and Kannada on app. *Times of India*. https://timesofindia.indiatimes.com/gadgets-news/flipkart-launches-support-for-tamil-telugu-and-kannada-on-app/articleshow/76558247.cms

[14] Flipkart. (2025, September 2). *Flipkart product returns process – your returns policy questions answered.* Flipkart Stories. Retrieved May 18, 2025, from https://stories.flipkart.com/flipkart-product-returns-2/

[15] Flipkart. (2023, February 9). *फ्लिपकार्ट प्रोडक्ट रिटर्न प्रक्रिया - रिटर्न पालिसी के सभी सवालों के जवाब.* Flipkart Stories. Retrieved May 18, 2025, from https://stories.flipkart.com/फ्लिपकार्ट-रिटर्न्स/

[16] Flipkart. (2023, February 9). *ஃபிளிப்கார்ட் தயாரிப்பு திரும்பப்பெறும் செயல்முறை – இது எவ்வாறு இயங்குகிறது மற்றும் நீங்கள் மனதில் கொள்ள வேண்டியவை.* Flipkart Stories. Retrieved May 18, 2025, from https://stories.flipkart.com/ஃப்ளிப்கார்ட்-திரும்புக/

[17] Flipkart. (2023, February 9). *ಫ್ಲಿಪ್‌ಕಾರ್ಟ್ ಉತ್ಪನ್ನ ಹಿಂದಿರುಗಿಸುವ ಪ್ರಕ್ರಿಯೆ - ಅದು ಹೇಗೆ ಕೆಲಸಮಾಡುತ್ತದೆ ಮತ್ತು ನೀವು ಏನನ್ನು ನೆನಪಿನಲ್ಲಿಡಬೇಕು.* Flipkart Stories. https://stories.flipkart.com/ಫ್ಲಿಪ್‌ಕಾರ್ಟ್-ಹಿಂದಿರುಗಿ/

[18] Microsoft. (2024). *Microsoft® Office Language Accessory Pack – Tamil*. Microsoft. Retrieved January 05, 2025, from https://www.microsoft.com/ta-in/download/details.aspx?id=51200

[19] Microsoft. (n.d.). *MSN | Personalised News, Top Headlines, Live Updates and more*. msn. Retrieved May 20, 2025, from https://www.msn.com/en-ae?ocid=msedgdhp&pc=U531&cvid=691c4d3821d94f51a3ac5e6d618a607e&ei=11

[20] Microsoft. (n.d.). *Microsoft translator | translate from English*. Microsoft Bing. Retrieved May 10, 2025, from https://www.bing.com/translator

[21] Microsoft. (2025, July 30). *Microsoft Change Locale*. Microsoft Services Agreement. Retrieved August 20, 2025, from https://www.microsoft.com/en-in/services agreement/locale

[22] Google. (n.d.). *Google*. Google. Retrieved May 20, 2025, from https://www.google.com/

[23] Google. (2024, May 22). *Google Terms of Service – Privacy & Terms – Google*. Google Policies. Retrieved August 20, 2025, from https://policies.google.com/terms?hl=en-IN&fg=1

[24] SketchEngine. (n.d.). *Setting up parallel and multilingual corpora*. SketchEngine. Retrieved October 23, 2025, from https://www.sketchengine.eu/guide/setting-up-parallel-corpora/#tab-id-2

[25] OpenAI. (n.d.). *ChatGPT*. [Large language model]. https://chatgpt.com/

[26] Google. (n.d.). *Welcome To Colab - Colab*. Colab. Retrieved October 10, 2025, from https://colab.research.google.com/

[27] YouTube. (2022, January 5). *Terms of Service*. YouTube IN. Retrieved May 20, 2025, from https://www.youtube.com/t/terms?hl=en&override_hl=1

[28] Apple Inc. (n.d.). *iPad User Guide*. Apple Support. Retrieved May 20, 2025, from https://support.apple.com/en-in/guide/ipad/welcome/ipados

[29] Apple Inc. (2025). *Find and download games in the Apple Games app on iPad*. iPad User Guide. Retrieved May 20, 2025, from https://support.apple.com/en-in/guide/ipad/ipad3aa36b02/ipados

[30] Apple Inc. (2025). *Add text on a Freeform board on iPad*. iPad User Guide. Retrieved May 20, 2025, from https://support.apple.com/en-in/guide/ipad/ipad5a22ec43/ipados

[31] Apple Inc. (2025). *iPad पर Freeform बोर्ड में टेक्स्ट जोड़ें. iPad यूज़र गाइड.* Retrieved May 20, 2025, from https://support.apple.com/hi-in/guide/ipad/ipad5a22ec43/ipados

[32] Apple Inc. (2025). *iPadನಲ್ಲಿನ Freeform ಬೋರ್ಡ್‌ನಲ್ಲಿ ಸ್ಟಿಕಿ ಟಿಪ್ಪಣಿಗಳು, ಆಕಾರಗಳು ಮತ್ತು ಪಠ್ಯ ಬಾಕ್ಸ್‌ಗಳಲ್ಲಿ ಪಠ್ಯವನ್ನು ಸೇರಿಸುವುದು. iPad ಬಳಕೆದಾರರ ಮಾರ್ಗದರ್ಶಿ.* Retrieved May 20, 2025, from https://support.apple.com/kn-in/guide/ipad/ipad5a22ec43/ipados

[33] Apple Inc. (2025). *iPadஇல் உள்ள Freeform போர்டில் ஸ்டிக்கி நோட்ஸ், வடிவங்கள்*

மற்றும் உரைப் பெட்டிகளில் உரையைச் சேர்த்தல். iPad பயனர் வழிகாட்டி. Retrieved May 20, 2025, from https://support.apple.com/ta-in/guide/ipad/ipad5a22ec43/ipados

[34] Lehmann, T. (1993). *A grammar of Modern Tamil* (2nd ed.). Pondicherry Institute of Linguistics and Culture.

[35] Apple Inc. (2025). *Wake, unlock, and lock iPad*. Apple Support. Retrieved May 21, 2025, from https://support.apple.com/en-in/guide/ipad/ipad9940ee8d/ipados

[36] Apple Inc. (2025). *iPad सक्रिय करें, अनलॉक और लॉक करें*. Retrieved May 21, 2025, from https://support.apple.com/hi-in/guide/ipad/ipad9940ee8d/ipados

[37] Apple Inc. (2025). *iPad ಅನ್ನು ಎಚ್ಚರಗೊಳಿಸಿ, ಅನ್‌ಲಾಕ್ ಮಾಡಿ ಮತ್ತು ಲಾಕ್ ಮಾಡಿ. iPad ಬಳಕೆದಾರರ ಮಾರ್ಗದರ್ಶಿ*. Retrieved May 21, 2025, from https://support.apple.com/kn-in/guide/ipad/ipad9940ee8d/ipados

[38] Apple Inc. (2025). *Send and reply to messages on iPad*. Apple Support. Retrieved May 21, 2025, from https://support.apple.com/en-in/guide/ipad/ipad99acb44a/ipados.

[39] Apple Inc. (2025). *iPadನಲ್ಲಿನ Freeform ಬೋರ್ಡ್‌ನಲ್ಲಿ ಸ್ಟಿಕಿ ಟಿಪ್ಪಣಿಗಳು, ಆಕಾರಗಳು ಮತ್ತು ಪಠ್ಯ ಬಾಕ್ಸ್‌ಗಳಲ್ಲಿ ಪಠ್ಯವನ್ನು ಸೇರಿಸುವುದು. iPad ಬಳಕೆದಾರರ ಮಾರ್ಗದರ್ಶಿ*. Retrieved May 21, 2025, from https://support.apple.com/kn-in/guide/ipad/ipad99acb44a/ipados.