International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

Towards Constructing Corpus of Punjabi N-grams Written in Gurmukhi Script

Er. Charanjiv Singh Saroa
Punjabi University Patiala, India
https://orcid.org/0000-0002-8940-9157

Dr. Kawaljeet Singh
Punjabi University Patiala, India
https://orcid.org/0000-0002-5314-0301

Abstract. The availability of a robust corpus is crucial for developing linguistic resources. For the Punjabi language, written in the Gurmukhi script, the scarcity of such a resource hinders the validation of various natural language processing (NLP) techniques. This paper addresses this gap by presenting the creation of a comprehensive corpus for Punjabi in Gurmukhi. The corpus, with approximately 23 million words drawn from diverse published materials, serves as a valuable foundation for NLP research. Additionally, the paper describes a dedicated corpus processing tool designed specifically for Punjabi. This tool employs a novel method for constructing word, bigram, and trigram levels of the corpus, applicable for building such resources for any script. As a demonstration, we showcase a generated dataset composed of approximately 15.5 million Punjabi words and 50 million characters

Keywords: NLP, Regional Languages, N-grams, UNICODE,

1 Introduction

Regional languages, steeped in the nuances of understanding, community, and culture, are essential components of our diverse linguistic landscape. However, in the contemporary era of Information and Communication Technology (ICT), these languages, whether spoken or written globally, often lag behind English due to a notable scarcity of resources, corpora, and language tools dedicated to regional languages. Unlike English, which enjoys a plethora of tools supporting grammar checking, spell checking, paraphrasing, and more, regional languages face a significant deficit in technological support.

The English language, as a global lingua franca, benefits from an abundance of datasets and corpora that fuel advancements in natural language processing (NLP) and related technologies. This linguistic asymmetry poses a challenge for regional languages, hindering their seamless integration into the digital sphere.

This research endeavors to address this disparity by focusing on the Punjabi language, a regional language spoken in the north Indian state of Punjab. This research aims to bridge this gap by creating a comprehensive corpus for Punjabi, serving as a critical foundation for developing language tools and propelling Punjabi into the digital future.

1.1 Regional Language

Communication, the act of giving, receiving, and exchanging information, is intrinsic to human interaction. Whether through verbal, non-verbal, or written means, language serves as the medium for effective communication. The vast linguistic diversity on our planet is challenging to quantify precisely. According to the Ethnologue (24th edition) (1), there are an estimated 7,139 living languages, with writing systems established for 4,065. The remaining languages are primarily oral, highlighting the dynamic nature of linguistic evolution influenced by culture, geography, and religion.

Regional languages are distinctive in that they are spoken in specific areas of a state or nation, ranging from small towns to larger regions. Their prevalence and characteristics vary according to the cultural, religious, and economic dynamics of each region. It's noteworthy that a nation may house hundreds of regional languages, each exhibiting diverse linguistic varieties. Importantly, a language transcends national borders and may be spoken across

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

multiple countries, underscoring the fluid and interconnected nature of linguistic communities. Table 1 highlights the top ten languages by number of native speakers.

Table 1: Top 10 Languages by number of native speakers (2)

Rank	Language	No of native speakers
1	Chinese	1.3 Billion
2	Spanish	471 Million
3	English	370 Million
4	Hindi	342 Million
5	Arabic	315 Million
6	Portuguese	232 Million
7	Bengali	229 Million
8	Russian	154 Million
9	Japanese	126 Million
10	Lahnda (Western bi)	118 Million

1.2 Script.

A script is a written representation of language that employs letters to symbolize consonants and vowels. These letters are combined to form syllables, which are then amalgamated to create words. Each word is separated by spaces to construct sentences. A variety of scripts is utilized to transcribe different languages worldwide. While scripts and languages are not inherently interconnected, most languages are predominantly written in a specific script. For instance, Hindi is primarily transcribed in the Devanagari script, and Punjabi is written in the Gurmukhi script. It is possible to write a language in a script other than its preferred one, as demonstrated by the sentence "वह जा रहा है," which is in the Hindi language and written in the Devanagari script. Conversely, the sentence "vah ja raha hai" is also in Hindi but transcribed in the Latin script, commonly known as the Roman script.

Table 2: The world's most popular writing scripts (3)

Rank	Name of script	Type	Population actively using (in millions)
1	Latin	Alphabet	over 4900
2	Chinese	Logographic	1340
3	Arabic	Abjad	660+
4	Devanagari	Abugida	608+
5	Bengali-Assamese	Abugida	300
6	Cyrillic	Alphabet	250
7	Kana	Syllabary	120
8	Javanese	Abugida	80
9	Hangul	Alphabet, featural	78.7
10	Telugu	Abugida	74
11	Tamil	Abugida	70
12	Gujarati	Abugida	48
13	Kannada	Abugida	45
14	Burmese	Abugida	39
15	Malayalam	Abugida	38
16	Thai	Abugida	38
17	Sundanese	Abugida	38
18	Gurmukhi	Abugida	22
19	Lao	Abugida	22
20	Odia	Abugida	21

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

1.3 Punjabi Language

According to the rajbhasha.gov.in website, 22 languages are included in the 8th schedule of the Indian Constitution (4), namely Punjabi, Hindi, Sanskrit, Santhali, Sindhi, Tamil, Assamese, Bengali, Bodo, Dogri, Gujarati, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Marathi, Maithili, Nepali, Oriya, Telugu, and Urdu. Punjabi is an Indo-Aryan language spoken by the majority of Punjabi people residing in both Pakistan and India. There are approximately 113 million native Punjabi speakers. According to the 2017 census of Pakistan, Punjabi is the most spoken first language in Pakistan, with 80.5 million native speakers. In India, it is the 11th most spoken language, with 31.1 million native speakers, as per the 2011 census. The language is also widely spoken by an international diaspora, particularly in Canada, the United States, and the United Kingdom.

1.4 Gurmukhi Script

The Gurmukhi script originated from the Landa (5) alphabet and was standardized in the 16th century by the 2nd Sikh Guru, Shri Guru Angad Dev Ji. The term "Gurmukhi" translates to "from the Guru's mouth." Primarily utilized for writing the Punjabi language, the script is horizontally written from left to right. Comprising 35 fundamental characters, the Gurmukhi script includes 10 vowels and modifiers, six additional modified consonants, and a total of 41 consonants, encompassing the 35 basic characters. Notably, there is no distinction between upper- and lower-case letters. Unlike the Greek and Roman alphabets, the Gurmukhi alphabet follows a logical organization, starting with vowels, followed by consonants (Gutturals, Palatals, Cerebral, Dentals, Labials), and semi-vowels (6). Table 3 illustrates the fundamental characters of the Gurmukhi Script.

Table 3: 35 Fundamental Characters of Gurmukhi Script

ੳ	ਅ	ੲ	ਸ	ਹ
ਕ	ਖ	ਗ	ਘ	ਙ
ਚ	ਛ	ਜ	ਝ	ਞ
ਟ	ਠ	ਡ	ਢ	ਣ
ਤ	ਥ	ਦ	ਧ	ਨ
ਪ	ਫ	ਬ	ਭ	ਮ
ਯ	ਰ	ਲ	ਵ	ੜ

Additional 6 Characters: ਸ਼ ਖ਼ ਗ਼ ਜ਼ ਫ਼ ਲ਼

Vowels and their associated modifiers: ਅ (none), ਆ (◌ਾ), ਇ (◌ਿ), ਈ (◌ੀ), ਉ (◌ੁ), ਊ (◌ੂ), ਏ (◌ੇ), ਐ (◌ੈ), ਓ (◌ੋ), ਔ (◌ੌ)

Other symbols: ◌ੰ (Tippi), ◌ਂ(Bindi), ◌ੱ(Adhak), ◌੍ (Halant)

2 Material and Methods

Before constructing a corpus, three major challenges must be addressed. First, what are the steps needed to generate the final corpus? Secondly, what is the objective of creating the corpus? And lastly, what are the design parameters for the corpus that guarantee its reliability? The steps for generating the Punjabi Corpus written in Gurmukhi text are depicted in Figure 1. In the initial step, among various sources for corpus collection, we opted for popular Punjabi books to gather words. This choice ensures the acquisition of authentic and validated Gurmukhi text. In the second stage, we amassed more than 17 million Punjabi words.

Our goal is to provide the Punjabi language research community with a corpus suitable for examining the occurrence of characters and words in a large dataset. Books serve as an excellent resource for this purpose due to their inclusion of authenticated content. The design of the corpus and the challenges encountered during its construction are discussed in the subsequent subsections.

Figure 1: Steps for Punjabi Words Corpus Generation

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

2.1.1 Corpus Design Criteria

Corpus design criteria play a crucial role in shaping the quality and effectiveness of linguistic corpora, which are extensive collections of texts or spoken language samples used for linguistic analysis. The complete Corpus design criteria that include language, script, sources of collection, technology, and tools used, mode of text, medium of text, year of text collected from books, etc. is shown in figure 2. The output of the corpus is 17 million words.

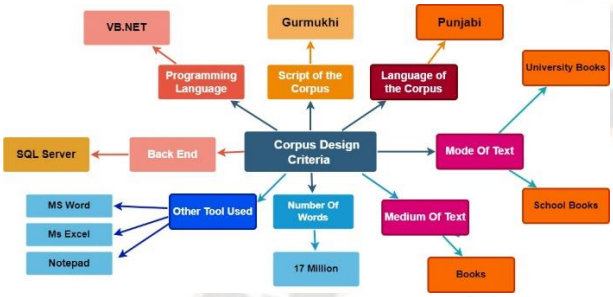

Figure 2: Corpus Design Criteria

2.2 Data Processing

There are hundreds of fonts available for almost every script. In most regional languages, when someone creates a font, they also change its keyboard layout (7). Due to this, information becomes useless with the change of font or unavailability of the desired font. This problem is mainly attributed to ASCII-based fonts. In Unicode, each character receives its unique code. To address this issue, we need to convert all the data available in ASCII-based fonts into Unicode-based fonts so that the information remains consistent regardless of the font used. This approach facilitates posting and searching for information on the web as well. In the data preparation phase, we convert the data into Unicode-based fonts. All the texts need to be normalized, including the removal of HTML tags, fixing punctuation, and eliminating text written in any other script.

2.3 Database Design

To create a corpus, we need to make seven tables in the database. Source Info table contains the information of the books, bookTBL is the input table that contains the data to be processed to create a corpus, and checkTBL contains information about the entries already processed or not. Wordfreqtbl is the first output table that includes all the words with their occurrence. This table is used to produce final tables with characters, bigram, and trigram with the frequency of occurrence and this information is stored in the characterTBL,bigramTBL, and trigramTBL.

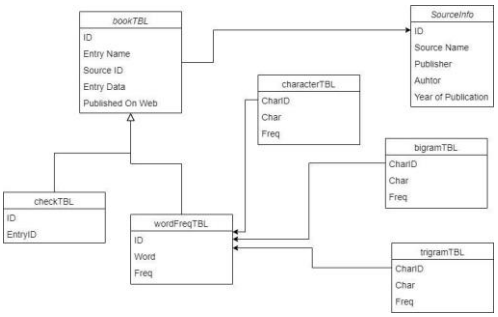

Figure 3. Shows the relationship between the tables used in the database

2.4 Method to create a corpus

The process of creating the corpus may seem easy and simple at first glance; however, this is not the case. We encountered various difficulties and challenges throughout the process. The initial challenge involved sourcing data in Unicode-based fonts. As the majority of the obtained data was in ASCII-based fonts, we had to develop an AI-based font converter (7) to facilitate the conversion of ASCII-based data into Unicode standard fonts this font converter detects the font and script automatically this font converter can be used as a model for other Regional languages also. Additionally, we had to perform data cleaning to ensure that the information used for processing adhered to the required script. For Each Book:

```
// Setting up database connection
SqlConnection connection = new
SqlConnection("ConnectionString");
SqlDataAdapter adp = new SqlDataAdapter();
adp.SelectCommand = new SqlCommand("SELECT *
FROM TableNAme WHERE BookID = 'book_id'",
connection);
// Creating DataSet and filling with data
DataSet ds = new DataSet();
adp.Fill(ds);
For Each Row in ds:
source = Row["source"];
entrydata = Row["entrydata"];
 // Checking if the value has already been checked
if not ccheckt(source, entrydata):
// Adding the value to the list of checked values
 addcheckt(source, entrydata);
// Processing data_entry
 data_entry = getRequiredDataEntry(entrydata);
 data_entry = removeHtmlTags(data_entry);
 data_entry = sclean(data_entry);
// Splitting the value into an array of words
words = splitIntoWords(data_entry);
For Each Word in words:
```

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: 25 July 2023 *Revised:* 12 September 2023 *Accepted:* 30 September 2023

```
                    // Checking if the word has already been
counted
                    if not checkword(Word):
                  // Adding the word to the list of counted
words
                    addword(Word);
                else:
                    // Adding the frequency of the word to the list
of counted words
                     addfreq(Word);
            End ForEach // Word
           // Calling functions to count and store character
frequencies
            Split the word in characters, bigrams & trigrams;
             Inschar(data_entry);
             Inschar2(data_entry);
             Inschar3(data_entry);
           End If // Value not checked
           End ForEach // Row
  End ForEach // Book
```

Data is stored in the entry form, and we use around a hundred thousand entries to create a corpus. As it is not possible to process all data at once, we create software that accepts entries range, and most of the time range is 3000 entries. After inputting the range, each entry was processed individually. Each entry is split into words, and then each word is stored in the database table as a new entry if the word is not in the table. Otherwise, the frequency of occurrence of the word is increased by one in the table. After processing all the entries, we get more than 17 million words and 2.5 million unique words. These words are used to create a corpus of characters, bigrams of characters, and trigrams of characters.

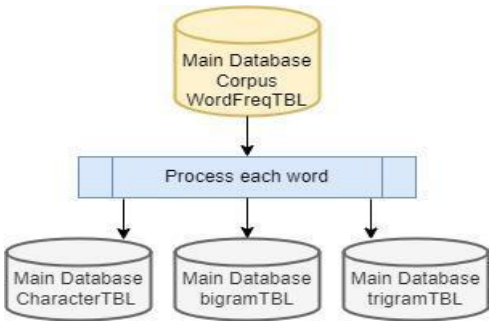

Figure 4. The figure shows the number of output tables that contain the required corpus.

3 Results & Discussion

The continuous sequence of n elements from a given sample of text or audio is defined as an N-gram. Depending on the application, the elements might be words, letters, or base pairs. N-grams are extracted from an available text or audio corpus (a large dataset of text) (8). An N-gram language model forecasts the likelihood of a particular N-gram occurring in any sequence of words or characters in the language. N-Grams can be unigram, bigram, trigram, etc. The output of our presented work is in the form of four corpora generated from the large dataset of the Punjabi language. These four corpora include a corpus of words, a corpus of characters, a corpus of bigrams, and a corpus of trigrams. All these corpora include the frequency of their occurrence in the data

Result 1: The first part of this process generates 1,56,66,836(approx. 15.5 million) words and 2,98,733 unique words. But there are some words from other scripts like Roman, Devanagari, etc after removing words from these scripts remaining unique words are 2,38,853, and the total number of words is 1,27,11,578(approximately 12.7 million).

Table 4. Top 15 words according to the occurrence in Punjabi data.

Rank	Word	Frequency
1	ਹੈ	424970
2	ਵਿਚ	386520
3	ਦੇ	368544
4	ਨੇ	336618
5	ਦੀ	276753
6	ਦਾ	262775
7	ਅਤੇ	255216
8	ਤੋਂ	247091
9	ਇਸ	220442
10	ਹਨ	149228
11	ਇਹ	119801
12	ਨਾਲ	101508
13	ਇਕ	101343
14	ਵੀ	88840
15	ਸੀ	88045

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
Article Received: 25 July 2023 Revised: 12 September 2023 Accepted: 30 September 2023

Result 2: In the second part of the process, find the frequency of occurrence of each character. In the data set of 1,27,11,578 words total no of characters found is 4,90,96,680(appx. 50 million) occurrence of ਾ(Kanna) is the maximum and it occurs 45,12,839 times.

Table 4. Top 15 characters according to the occurrence in Punjabi data.

Rank	Word	Frequency
1	ਾ	4512839
2	ਰ	2934160
3	ਿ	2499080
4	ੀ	2389640
5	ੇ	2339404
6	ਦ	2251033
7	ਨ	2035609
8	ਹ	2032734
9	ਸ	2016493
10	ਤ	1948564
11	ਕ	1901233
12	ਂ	1628182
13	ਵ	1557848
14	ਲ	1407146
15	ਪ	1115761

Result 3: In this part of the process, bigrams are generated. From the data set total no of Bigrams is 2,51,97,505(appx. 25million). Among them, 3176 are unique bigrams that are found in the data. The occurrence of ਦੇ (de) is maximum, and it occurs 5,67,061 times.

Table 5. Top 15 bigram according to the occurrence in Punjabi data.

Rank	Word	Frequency
1	ਦੇ	567061
2	ਦਾ	556488
3	ਦੀ	500157
4	ਹੈ	438643
5	ਨਾ	385216
6	ਨੇ	372966
7	ਰਾ	366554
8	ਤੇ	334970
9	ਜਾ	332090
10	ਆਂ	300438
11	ਵਾ	292957
12	ਤਾ	288461
13	ਕਾ	284655
14	ਤੋ	269062
15	ਹਾ	267638

Result 4: In the last part of the process, trigrams are generated. The total no of trigram founds is 1,67,33,660(approx. 16 million). Among them, 29387 are unique trigrams that are found in the data. The occurrence of ਵਿਚ(vich) is maximum, and it occurs 4,10,739 times.

Table 6. Top 15 trigram according to the occurrence in Punjabi data.

Rank	Word	Frequency
1	ਵਿਚ	410739
2	ਤੋਂ	260962
3	ਅਤੇ	255460
4	ਜਾਂ	169767
5	ਪ੍	167476
6	ਕਾਰ	110721
7	ਹਾਂ	109266
8	ਨਾਲ	106871
9	ਕੀਤ	97075
10	ਨਾਂ	89674
11	ਰਿਆ	86272
12	ਹੁੰ	86092
13	ਦੀਆ	75214
14	ਨ੍ਹ	70157
15	ਤਾਂ	66777

4 Conclusion and Future Work

In conclusion, this paper centers on the generation of a corpus for a regional language, specifically Punjabi words written in the Gurmukhi script. Following the corpus generation, a frequency analysis of Punjabi words was conducted to identify the most commonly used words and characters. The study revealed that ਾ(Kanna) is the most frequently occurring character, and ਹੈ is the most frequently written word in Punjabi text. Additionally, the identification of the most frequent trigrams and bigrams opens avenues for the development of various conversion tools, such as Punjabi text to Braille script, tools for semantic analysis of Punjabi language words, and applications like spell checkers and text-to-speech generation. The generated corpus can be used to create effective spell checkers, translation systems, text-to-speech systems, text predictions, etc. So generated corpus holds significant value for the research community, particularly those engaged in natural language processing within the Punjabi language domain.

References

1. Ehnologue (2022) The most authoritative source on the languages of the world Available in : https://www.ethnologue.com/statistics/ Accessed on: September 24, 2022.
2. Lane James (2023) The-10-most-spoken-languages-in-the-world. Available in: https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world Accessed on: February 20, 2023
3. AmberP (2019) The World's Most Popular Writing Scripts: Available in: https://www.worldatlas.com/articles/the-world-s-most-popular-writing-scripts.html Accessed on: September 25, 2022.
4. Department of Official Language, Government of India (2022) Available in https://rajbhasha.gov.in/en/languages-included-eighth-schedule-indian-constitution Accessed on: September 25, 2022.
5. Singh, Harjeet, Khanna, Ravinder and Goyal, Vishal (2013) Comparative Study of Standard Punjabi and Malawi Dialect with regard to Machine Translation, International Journal of Engineering Sciences 8:2229-6913.
6. Siddharth, Kartar Singh,JangidMahesh, Dhir Renu, Rani Rajneesh (2011) Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features, International Journal of Computer Science & Engineering 3: 2332-2345.
7. Singh, Rajwinder and Saroa, Charanjiv Singh (2017) Multilingual Conversation ASCII to Unicode, Conference proceedings of sixth International Conference on Information Technology Convergence and Services, AIRCC Publishing Corporation, pp 83-94.
8. Pawan, Soda Kashish, Arora Simran (2022) N-Gram Language Modelling with NLTK Available

inhttps://www.geeksforgeeks.org/n-gram-language-modelling-with-nltk Accessed on : September 25, 2022.
9. Braj B. Kachru, University of Illinois, Chicago, Yamuna Kachru (2008) Language in South Asia. Cambridge University Press, New York.
10. UNESCO (2022) A decade to prevent the disappearance of 3,000 languages Available in https://www.iesalc.unesco.org/en/2022/02/21/a-decade-to-prevent-the-disappearance-of-3000-languages Accessed on September 24, 2022.