Towards Sindhi Corpus Construction

Mutee U Rahman
Department of Computer Science, Isra University, Hyderabad Sindh 71000, Pakistan
muteeurahman@gmail.com

Abstract

The paper discusses the current state of Sindhi corpus construction in detail. Sindhi corpus development issues including corpus acquisition, preprocessing, and tokenization are discussed in detail. Preliminary results and observations which include letter unigram, bigram and trigram frequencies; word frequencies and word bigram frequencies are presented. Current state of Sindhi corpus with its limitations and future work is also discussed. The paper also explores the orthography and script of Sindhi language with reference to corpus development.

1. Introduction

Sindhi is one of the major languages of Pakistan spoken by approximately 30-40 million people [1][2]. Sindhi is being frequently used on internet. Sindhi blogs, literary websites, online newspapers and discussion forums are increasing day by day. After Urdu Sindhi is the second largest written language of Pakistan. Despite of its online usage and popularity only few language processing resources are available for NLP researchers which include lexicon, fonts and simple word processors. The development of Sindhi language processing resources like linguistic corpora and comprehensive computational lexicon are not even initiated.

Sindhi is being written in Persio-Arabic (سنڌي) , Devnagri (सिन्धी) and roman (sindhi) scripts. Persio-Arabic script is most common script for Sindhi writings in Pakistan and India. Devnagri script is also being used for Sindhi writing in India. Roman script (though not yet standardized) is also getting popularity. Very few written documents are available in roman script but it is being used frequently for communications on internet and cell phones and other smart devices. Due to the fact that most of the online and offline written material of Sindhi is available in Persio-Arabic script Sindhi corpus being constructed is in Persio-Arabic script using UTF-16 encoding.

Following sections discuss the existing work in Pakistani language corpora, orthography and script of Sindhi Language, corpus construction issues, corpus acquisition, preprocessing, tokenization and results of preliminary statistical analysis. Finally the future work is discussed along-with conclusion.

2. Previous work

Apart from fonts, keyboard design [3] and few digital dictionaries [4] Sindhi language processing resources are not available publically. Studies or development projects for resources like linguistic corpora and comprehensive computational lexicon are not even initiated. Various research organizations and individuals are working for the development of linguistic corpora of different Pakistani languages. For Urdu EMILLE [5], Baker Riaz corpus [6], jang newspaper corpus [7], and parallel English Urdu and Nepali corpus [8] are some key examples. For Pashto the projects include BBN Byblos Pashto OCR System [9] and Machine readable Pashto text corpus being developed at University of Peshawar [10]. The first Punjabi language corpus was developed by Central Institute of Indian Languages (CIIL) India [11]. Hindi and Punjabi parallel corpus developed by CDAC Noida is another useful linguistic corpora available. One cannot find such type of linguistic corpora for Sindhi, Balouchi, Siraiki and many other Pakistani languages. In contrast to other Pakistani languages (Excluding Urdu) Sindhi text in electronic format is easily available and is being continuously collected for corpus under discussion.

3. Orthography and script of Sindhi language

Sindhi is written in Persio-Arabic script based on extended Arabic character set in Naskh style. Sindhi

alphabet is comprised of 52 letters shown in figure 1. The alphabet contains basic letters like ب، ب، ا and secondary letters like جھ، and گھ which are aspirated versions of ج and گ.

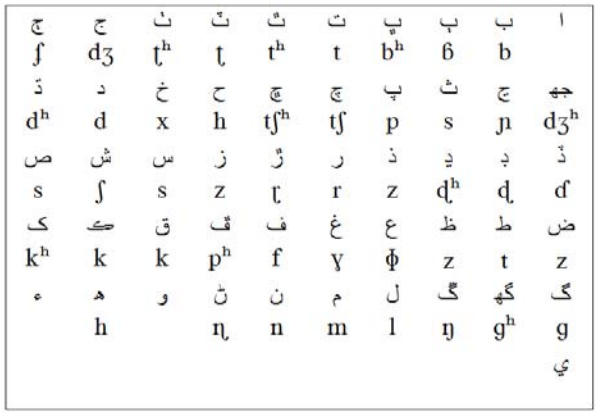

Figure 1. Sindhi alphabet.

Sindhi words always end in a vowel [12]; this vocalic ending is optionally marked by diacritics in written text. Diacritics are also used inside words to represent additional vocalic features. Absence of diacritics in written text sometimes cause semantic ambiguities. For instance the word ڊَهڻ (to push) and ڊُهڻ (bog) are semantically ambiguous without diacritics. Diacritics used in Sindhi are shown in Figure 2.

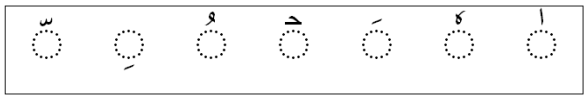

Figure 2. Diacritics used in Sindhi.

Sindhi has its own numerals based on Persio-Arabic numerals shown in figure 3. Use of Hindu-Arabic numerals is also very common in Sindhi writings. Special symbols shown in figure 3 are also used in Sindhi written text.

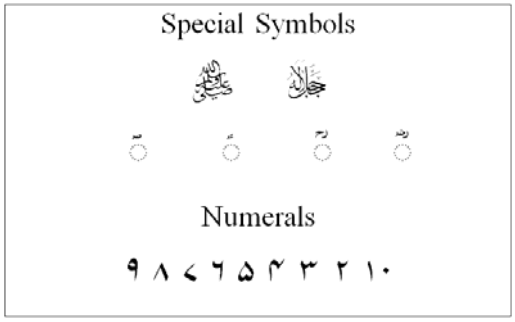

Figure 3. Special symbols and Numerals used in Sindhi written text.

4. Sindhi Corpus Development

After Unicode support and Unicode based Sindhi keyboard design [13] availability of Unicode based Sindhi text on Internet is increasing day by day. Key factor behind the motivation of Sindhi corpus construction is availability of online text in Sindhi newspapers, blogs, literary websites and discussion forums. Despite of the fact that available online resources do not provide huge amount of text but they are increasing day by day and corpus is being collected continuously. Software routines for preprocessing, normalization, tokenization and frequency calculation are implemented in C# using Microsoft .net framework libraries.

4.1. Corpus Acquisition

Data is gathered from various domains which include news, blogs, literature, essays, and letters. Different subdomains include current affairs, sports, showbiz, short stories, discussions and opinions. Sources of data collection are shown in Table 2.

Table 1. Sources of data collection.

Source	URL(s)
Daily Kawish	http://www.thekawish.com
Daily Awami Awaz	http://www.awamiawaz.com
Daily Ibrat	http://dailyibrat.com
Blogs	http://shikarpuri.wordpress.com
Literary Writings	http://voiceofsindh.net http://sindhsalamat.com

4.2. Preprocessing and Normalization

Almost all data gathered was already in Unicode format but nevertheless all the collected text is converted into standard UTF-16 encoding. Letters represented by multiple Unicode points and equivalent representations of composed and decomposed form [14] are reduced to same underlying form. Letters with aspirated versions like گھ which are combinations of two Unicode characters (for instance گ and ھ in case of گھ) are considered single letters while dealing with text processing.

4.3. Tokenization

For tokenization white spaces, punctuation markers, special symbols (like $, %, # etc.) and digits are used as word boundaries. White space word boundary consideration caused problem of embedded space word breaking (For example the single word صاحب قدرت is divided into two words صاحب and قدرت) is tackled out by using the same technique used for Urdu [15]. Another problem in Sindhi word tokenization occurs when two special words ۾ (in) and ۽ (and) occurred without space like ملائڻ۾ (meː mila:iɳa) and this was tokenized as a single word. Also in case ڪتاب۽قلم kita:ba ain qalama (book and pen) in which three words without space are there and were tokenized as single word. Same problem was observed with all the words with non-connective ending like ڪيرپي k^{h}i:ra pi:a (drink milk) or starting letters سنڌاندر $sind^{h}$a ander (in Sindh). Semiautomatic (software based + manual) approach was used to overcome this problem.

5. Results and observations

A total of 4.1 million word corpus analyzed quantitatively. This preliminary analysis includes letter frequency analysis, letter bigram analysis, letter trigram analysis, word frequency analysis, and word bigram analysis. These quantitative results are discussed in following sections.

5.1. Letter frequencies

A total of 13,968,112 characters in the corpus were analyzed while calculating letter frequencies. Along-with 52 letters of Sindhi alphabet آ was also considered as a single letter because of its use in Sindhi keyboard as a single letter and single Unicode representation. It was observed that most frequently occurred letter was vowel ي while least frequently occurred letter was consonant ڳ. Table 2 shows top 20 most frequently occurred letters in Sindhi corpus with their percentage.

While analyzing frequencies it was observed that frequency distribution of individual letters in single file of 50,000 or more words was identical to the letter frequency distribution of whole corpus. This similarity can be seen in graphs of figure 3 and 4.

Letter bigram and trigram frequencies were also analyzed. It can be seen that almost 50% of top 20 most frequent bigrams are valid two letter words like ان, جي, هن and ڪي. Same is the case with trigrams where this ratio is more than 60%. Top 20 most frequent bigram and trigram percentages are shown in Tables 3 and 4 respectively.

Table 2.Top 20 most frequent letters.

S.No.	letter	Percent	S.No.	Letter	Percent
1	ي	13.77%	11	ڪ	3.25%
2	ا	11.42%	12	س	3.23%
3	ن	8.99%	13	د	2.50%
4	و	7.84%	14	ب	2.00%
5	ه	6.27%	15	پ	1.80%
6	ر	6.15%	16	آ	1.18%
7	م	3.73%	17	ٽ	1.16%
8	ج	3.64%	18	ک	1.16%
9	ل	3.30%	19	ع	0.99%
10	ت	3.26%	20	ڻ	0.94%

Table 3. Top 20 bigrams in Sindhi corpus.

S.No.	Bigram	Percent	S.No.	Bigram	Percent
1	ان	3.16%	11	ون	1.18%
2	جي	2.55%	12	يا	1.10%
3	ري	1.95%	13	آه	1.10%
4	هن	1.80%	14	جو	1.07%
5	يو	1.79%	15	وا	1.02%
6	ار	1.79%	16	ال	1.01%
7	هي	1.79%	17	ڪي	0.99%
8	ين	1.69%	18	ور	0.97%
9	نه	1.28%	19	لا	0.95%
10	ند	1.27%	20	تي	0.93%

Table 4. Top 20 letter trigrams in Sindhi Corpus.

S.No.	Trigram	Percent	S.No.	Trigram	Percent
1	آهي	1.40%	11	هنج	0.45%
2	نهن	1.34%	12	آهن	0.44%
3	اري	0.81%	13	ڪيو	0.44%
4	يون	0.74%	14	انه	0.42%
5	ڪري	0.71%	15	ندو	0.41%
6	کان	0.61%	16	اٿي	0.40%
7	يند	0.60%	17	نجي	0.36%
8	ندي	0.53%	18	ڏهن	0.35%
9	وار	0.47%	19	پنه	0.35%
10	مان	0.46%	20	دار	0.35%

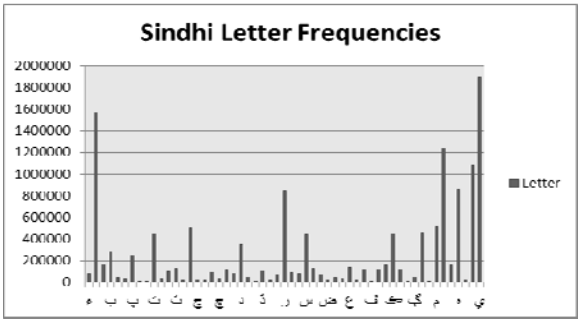

Figure 4. Letter frequency distribution in Sindhi corpus.

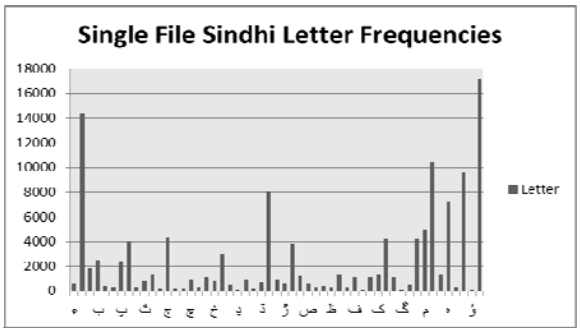

Figure 5. Letter frequency distribution in a single file.

5.2. Word frequencies

Total of 4.1 million words were analyzed and 70,576 distinct word forms were found. Most frequently occurring words include case markers (like ۾, تي and کان) and auxiliary/incomplete verbs (like آهي and آهن). Postposition جي has highest frequency of occurrence as shown in Table 5.

Table 5. Top 20 most frequent words in Sindhi corpus.

S.No.	word	Percent	S.No.	word	Percent
1	جي	3.71%	11	ڪري	0.69%
2	۾	2.44%	12	سان	0.69%
3	ءِ	2.17%	13	ان	0.67%
4	ته	1.78%	14	کان	0.63%
5	آهي	1.61%	15	ٿي	0.57%
6	کي	1.61%	16	آهن	0.55%
7	جو	1.50%	17	لاءِ	0.51%
8	تي	1.05%	18	هن	0.50%
9	به	0.82%	19	هو	0.50%
10	نه	0.71%	20	ڪيو	0.46%

Word bigram occurrences are also calculated and are shown in Table 6. The proper name bigram بينظير ڀٽو is among the top 10 bigrams. This is because of the current affairs domain contains essays and newspaper columns about the life of former prime minister Benazir Bhutto.

6. Future work

Corpus is being continuously collected and results are being updated. Currently corpus is simply a UTF-16 encoded text collection. Study are in progress for proper annotations, POS tagging, corpus based lexicon development and n-gram based text categorization. Sindhi tokenization algorithm need to be worked out for the problems discussed in section 4.3. Due to absence of standard sentence termination punctuation marker in Sindhi; full stop comma and other punctuation markers are used as sentence terminators in Sindhi text writings. Sentence segmentation is another key area to be worked out. More specific Sindhi computational linguistic studies are needed for further development and maturity of corpus. For example currently there is no comprehensive POS tagging algorithm available for Sindhi. Presently available POS tagging algorithm for Sindhi [16] need to be analyzed and extended further. Sindhi tagset need to be designed before POS tagging of the corpus. Qualitative, quantitative improvements, proper annotations and comprehensive statistical analysis are areas to be extensively worked out.

Table 6. Top 10 most frequent word bigrams.

S.No.	Word bigram	Percentage
1	چيو ته	7.52
2	آهي ته	6.75
3	هن جي	2.66
4	بينظير ڀٽو	1.93
5	سنڌ جي	1.84
6	ان جي	1.72
7	جڏهن ته	1.60
8	هن چيو	1.60
9	ڪيو ويو	1.44
10	ويو آهي	1.21

7. Conclusion

In absence of language processing resources of Sindhi language Sindhi corpus construction project is a valuable initiative. Regardless of its size and preliminary results the corpus in its current state will provide basis for further natural language processing studies of Sindhi language. Letter frequencies including bigram and trigram frequencies provide basis for intelligent text processing and compact keyboard design for cell phones and other smart devices. Word level unigram and bigram frequencies provide basis for spelling corrections and automatic sentence completion applications. Further developments in corpus will be useful for advanced language processing tasks like morphological analysis, syntax analysis, semantic analysis, information retrieval and extraction and machine translation.

8. References

[1] Sindhi Language Authority. Official Website. http://www.sindhila.org. (Accessed 2010).

[2] C. Jennier. (2006), The Sindhi Language. In K. Brown (ed). *Encyclopedia of Language and Linguistics, 2nd Edition*, v. 11:384-386. Oxford: Elsevier.

[3] Bhurgri A.M. 2010 . A Breakthrough in use of Sindhi on Internet *Indus Asia Online Journal* http://iaoj.wordpress.com/2010/07/07/a-breakthrough-in-use-of-sindhi-on-internet/ (Accessed 2010).

[4] Sindhi English Dictionary. http://www.crulp.org/sed/ (Accessed 2010).

[5] A.M. McEnery, P. Baker, R. Gaizauskas, and H. Cunningham. EMILLE: Building a Corpus of South Asian Languages. *Vivek, A Quarterly in Artificial Intelligence*, 2000, 13(3):23–32.

[6] D. Becker, K. Riaz. A Study in Urdu Corpus Construction. Proceedings of the *3rd Workshop on Asian Language Resources and International Standardization at the 19th International Conference on Computational Linguistics*. August 2002.

[7] Hussain, S. Resources for Urdu Language Processing. *The Proceedings of the 6th Workshop on Asian Language Resources, IJCNLP'08*, IIIT Hyderabad, India. 2008.

[8] Urdu, Nepali and English Parallel Corpus, CRULP http://crulp.org/software/ling_resources/UrduNepaliEnglishP-arallelCorpus.htm (Accessed: 2010).

[9] Debro, M. et al. The BBN Byblos Pashto OCR System. ACM Press. 2004.

[10] M. A. Khan and F. T. Zuhra, "A General-Purpose Monitor Corpus of Written Pashto", in proc. *Conference on Corpus Linguistics*, Birmingham, July, 2007

[11] G S Lehal, "A Survey of the State of the Art in Punjabi Language Processing", *Language In India, Volume 9*, No. 10, pp. 9-23 2009.

[12] Rahman. M. Sindhi Morphology and Noun Inflections. In the proceedings of *Conference on Language and Technology CLT09*. Crulp Lahore. 2009.

[13] http://www.bhurgri.com/ Accessed 2010

[14] H. Sarmad & Durrani N. Sindhi in PAN Localization A Study on Collation of Languages from Developing Asia 2008.

[15] Ijaz, M. and Hussain, S. (2007). Corpus Based Urdu Lexicon Development. In the *Proceedings of Conference on Language Technology '07,* University of Peshawar, Peshawar, Pakistan.

[16] Mahar J.A., Memon G.Q. "Rule Based Part of Speech Tagging of Sindhi Language," icsap, *International Conference on Signal Acquisition and Processing*, pp.101-106, 2010.