TO BUILD CORPUS OF SINDHI LANGUAGE

Fida Hussain Khoso
Dawood University of Engineering & Technology Karachi. Indus University. Karachi (Pakistan)
E–mail: fidahussain.khoso@duet.edu.pk

Mashooque Ahmed Memon
Benazir Bhutto Shaheed University Lyari. Karachi (Pakistan)
E–mail: pashamorai786@gmail.com

Haque Nawaz
Sindh Madressatul Islam University. Karachi (Pakistan)
E–mail: hnlashari@smiu.edu.pk

Sayed Hyder Abbas Musavi
Indus University. Karachi (Pakistan)
E–mail: dean@indus.edu.pk

Recepción: 05/03/2019 **Aceptación:** 21/03/2019 **Publicación:** 17/05/2019

Citación sugerida:
Khoso, F. H., Memon, M. A., Nawaz, H. y Abbas Musavi, S. H. (2019). To build corpus of Sindhi language. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Edición Especial, Mayo 2019*, pp. 100–115. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100–115

Suggested citation:
Khoso, F. H., Memon, M. A., Nawaz, H. & Abbas Musavi, S. H. (2019). To build corpus of Sindhi language. *3C Tecnología. Glosas de innovación aplicadas a la pyme. Special Issue, May 2019*, pp. 100–115. doi: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100–115

ABSTRACT

The present day state of Sindhi corpus construction is elaborated in detail in this paper. The issues like corpus acquisition, tokenization and preprocessing have been analyzed and discussed minutely for Sindhi corpus enhancement. Initial observations and results are included for letter unigram, bigram and trigram frequencies. There has been discussed the present status of Sindhi corpus in perspective of restriction and future work. Orthography and script were also explored in this paper with reference to corpus development. Basically the word corpus was used first time by German Scholar (Das Corpus). The plural of corpus is corpora, which is used for huge text data consists of millions and billions of text data. The task of Natural Language Processing was very challenging because there was the scarcity of resources for computational linguistics and research. Different text corpora have been made in different languages of different countries, after reviewing the corpora of different languages of various countries, we are trying to make the corpus for Sindhi language.

KEYWORDS

We NLP, Corpora, Linguistic, Lexicon, Phoneme.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100-115

1. INTRODUCTION

About thirty to forty million people of Pakistan speak Sindhi language and it is a big language. On internet Sindhi language is vastly used. The number of news papers literary websites and blogs of Sindhi language is increasing daily. The lexicon, fonts and common words processes are included and available for NLP researchers and this is the evidence of usage and popularity of online. In Sindhi language such as linguistic corpora are not initiated for the enhancement of Sindhi language processing resources.

Sindhi language is being used and written in Arabic-Persian, Devanagari and Roman letters. For Sindhi language in India Devanagari letters are also used. Same as the Roman script is getting popularity for Sindhi language. On smart phone devices, cell phones and communications on internet have been used and available in Roman script for very few documents. It is unfortunate that the linguistic corpora and detailed computational lexicon are still not initiated because it was very essential for the development of Sindhi language processing resources. It is factual position that in Sindhi language that excess written material is available for offline and online. Sindhi Corpus the script is Persio-Arabic which has been built in Persio-Arabic script using UTF-16 in coding. In these sections we are discussing the orthography and Sindhi language corpus script which is achieved are results of initial statistical analysis, preprocessing the issues of corpus construction of Pakistani language corpora. In this conclusion we have finally discussed the future work (Mahar & Memon, 2010).

2. PREVIOUS WORK

As for the Sindhi language processing resources concerned, apart from few digital dictionaries, key board design and fonts, these are not generally and publically available. Even in Sindhi language for resources like comprehensive computational lexicon and linguistic corpora, studies or development projects are not even initiated. Because of the improvement of linguistic corpus of various languages of Pakistan the different research organizations and individuals are

working. We can give the example of Jang newspaper corpus (Hussain & Durrani, 2008; Becker & Riaz, 2002) .

At the university of Peshawar machine readable Pashto corpus is being developed and the other project BBN Byblos Pashto OCR system is included. The central institute of Indian languages (CIIL) of India had developed first time the Punjabi language corpus. CDAC-Noida which is another useful linguistic corpora has developed Hindi and Punjabi parallel corpus. There are not available such kind of linguistic corpora for many Pakistani languages such as Siraiki, Balochi and Sindhi. It is contrary that Sindhi text is easily available in electronic format and the corpus under discussion is being collected ceaselessly where as Urdu does not possess this facility.

3. SINDHI LANGUAGE HANDWRITING

In Naskh style which is base on elaborated Arabic character, Sindhi is written in Persio-Arabic script. Sindhi letters are 52 in numbers as shown in Figure 1. The basic letters are contains in alphabet like ١,پ,ٻ and other letters for secondary just like ڄ and ڏ which are used in Sindhi language.

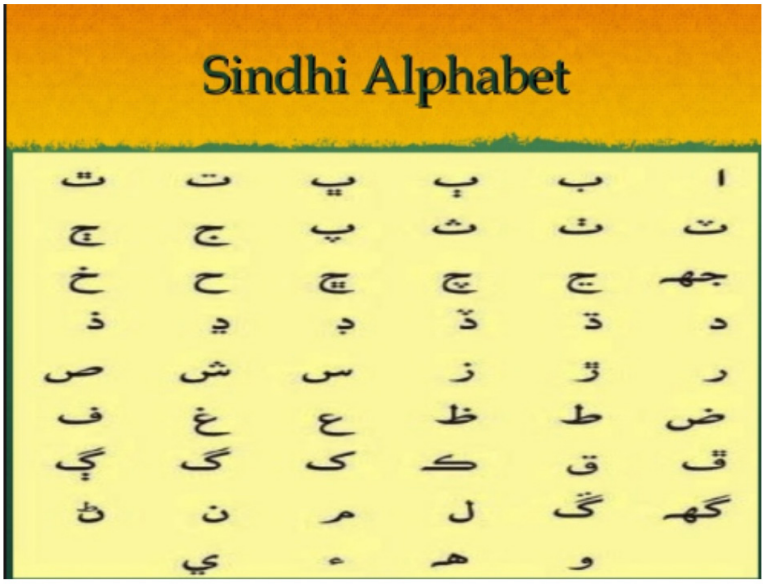

Figure 1. Sindhi Alphabet.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100-115

The words of Sindhi language are constantly ended with vowels. Diacritics in written text optionally Mark this vocalic ending. To represent additional voice features, the diacritics are also use. Sometimes semantic ambiguities are caused by the absence of diacritic in written text.

Having consists of Persio-Arabic digits which are appears in graph 2; Sindhi language has its own numerals. In Sindhi writing numerals are extremely common usage for Hindi-Arabic. In Figure 2 particular symbols are also used.

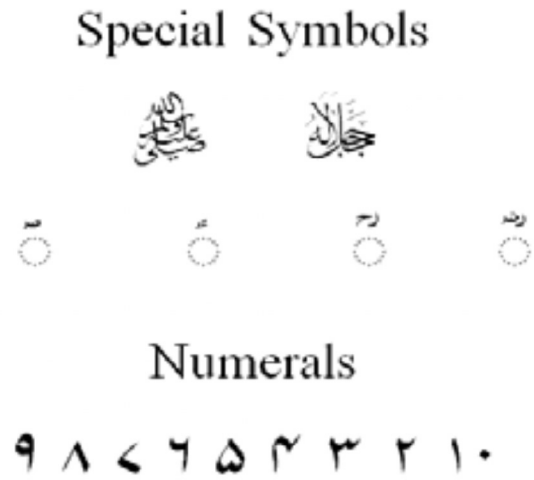

Figure 2. Numerals are used in written text of Sindhi language.

4. THE PROGRESS OF SINDHI LANGUAGE TEXT CORPUS

It is obvious that accessible resources upon internet do not provide huge amount of Sindhi text data. On daily basis for Unicode based Sindhi text on internet is enhancing very fast after Sindhi keyboard based on Unicode and Unicode support. Include e-mail addresses if possible. Follow the author information by two blank lines before main text. The accessibility of Sindhi corpus construction of newspapers, blogs, discussions forums and literary websites is the key factor to motivate Sindhi corpus construction. When we consider the importance of text corpus of Sindhi language along with other NLP improvements and linguistics Sindhi text corpus is fabricated. The corpus of Sindhi is being achieved constantly and the vast amount was not provided by the resource of online. In

"C" utilizing microsoft.net framework libraries the preprocessing tokenization, frequency calculation and normalization are implemented in software routines. To develop the text corpus based Sindhi sentimental analysis, language variation and sentiment analysis of aspect based and for other future research.

4.1. CORPUS ACQUISITION

From various domains which include letters, essays, literature, news and blogs the basic information is collected. Current affairs, short stories, sports, showbiz, opinions and discussions are included in different sub domains.

Table 1. Data collection sources.

Source	Web sites
Jhoongar	www.dailyjhoongar.com
Kawish	www.thekawish.com
Ibrat	www.dailyibrat.com
NLP	www.nlp.com
Sindhi virtual library	www.library.sindhila.org
Awami Awaz	www.awamiawaz.com

4.2. SUBSTANCES AND PROCEDURES

By utilizing the process techniques of text corpus building, the text corpus progress is done. From online blogs, websites, books and newspapers of Sindhi language, the text is achieved. The morphological analysis, Sentiment analysis, stemming, lemmatization, sentiment analysis of aspect based and tagging are the parts of the speech and for tokenization the text corpus of Sindhi is processed.

4.3. NORMALIZATION AND PREPROCESSING

Although all the gathered text has been converted into standard UTF-16 in coding the overall data which was collected and available in Unicode format.

An equivalent representation of data & information together are reduced to same underline form and they are represented by letters. The combination of two Unicode characters are aspirated versions گھ for instance گ and ھ when dealing with text processing they are considered as single letters.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100-115

4.4. HOW DO THE WORDS WORK IN SINDHI AND THEIR IDENTIFICATION?

Words are the identification of experiment and experiences of human being. What we do observe, listen, feel, testify and other actions all of these things are dependent on our thinking, conception and experiences. One who talks or writes tells the same words according to his perspective and assessment. Considering all these things there are some words in the following. Although the same word is being presented in different meanings so that it will clearly be understood and assessed the exits meaning of the word on narration. By the use of machine in proper way, the suffixes and affixes can be removed from inflected text in Sindhi (Rahman, 2009).

Carrying out the analysis of Sindhi text, the discussion on text corpus is very suitable.

Table 2. Sentiments and identified of Sindhi text corpus.

Sindhi words	English Meaning	Usage in English	Sentiments or Usage in Sindhi
مڙس	Brave	No doubt Dodo is brave	دودو بيشڪ **مڙس** آهي. (بهادر)
بالغ	Adult	Earlier Arshid was a child but now he is adult	اختر اڳ ۾ ڇوڪر هو پر هاڻي **مڙس** ٿيو آهي
شوهر	Husband	Dawood is husband of Zeenat	دائود زينت جو **مڙس** آهي(شوهر)
ماڻهو	Man	Who is he	هي ڪير **مڙس** آهي
مرد	Man	The greatness of a person is in keeping promise.	مڙس جو شان آهي ت هو پنهنجي واعدي تي قائم رهي

In data mining application and research, text analysis is an important topic because the scientific text and analysis of educational political and social text are internet resources and they produce large stuff of text. The useful data and information are extracted by organizations to analyze the text corpus therefore it becomes easier to translate the language and for decision makers to take good decisions. The feature distribution and language variation can be observed for the task of information retrieval.

4.5. TOKENIZATION

As $, %, #, etc along with digits are used as word boundaries and for tokenization they are white spaces punctuation markers and special symbols. The problem of embedded space world breaking is called by white space word boundary consideration. For instance any one word is bifurcated in two words نالو and منهجو and by using the same technique for Urdu the problem be resolved (Ijaz & Hussain, 2007).

If we do compare two words which are special ۾ (in) and ۽ (and) are occurred, another problem in Sindhi word tokenization appears ملاءڻ .(milana) and this was tokenize a single work. An example is here that قلم ۽ڪاپي (pen and note book) and these 03 words sans gap are here by tokenized as single word. In Sindh there are the same problems with all the words which have non connective ending.

5. OBSERVATIONS AND RESULTS

In numbers the whole word corpus of 4.1 million have been analyzed. The letter frequency analyzed, letter trigram analysis, analysis of letter bigram, word bigram analysis and word frequency analysis are included in this basic analysis.

5.1. MECHANISM UNDERSTANDABLE CORPUS

The languages of the world can be understood by people with the help of computational technology advancement. In this connection the role of linguists and computational linguistics is very important. It is necessary that the text corpus must be in machine readable form. To read and recognized the Sindhi text corpus through machine, Unicode utp-8 is used. On the basis of polarity analysis the sentiment analysis the sentiment analysis has been carried out. The sentiment analysis of text corpus document is shown by the results and it presents the features of outputs along with opinion and sentiment of each feature independently.

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100-115

5.2. FREQUENCIES OF LETTER

When calculating frequencies of letter the aggregate number of 139,886,112 characters was analyzed of corpus. The letter 'آ' we as well seen a single letter out of 52 letters of Sindhi alphabets and the reason is it is used a single letter in Sindhi keyboard and single Unicode representation. In Sindhi the least frequently occurred letter was consonant ڳ and the most frequently occurred letter was vowel ئ.

Top 20 a most frequently occurred letters with their percentage in Sindhi database are shown in Table 3.

Table 3. Twenty letters for frequent.

S.No.	Letter	Percent	S.No.	Letter	Percent
1	غ	12.25%	11	س	3.25%
2	ق	11.62%	12	ڪ	3.27%
3	و	7.84%	13	د	2.50%
4	ن	8.99%	14	ب	2.00%
5	ر	6.16%	15	ڀ	1.80%
6	ه	6.26%	16	آ	1.18%
7	م	3.73%	17	ٽ	1.17%
8	ج	3.64%	18	ک	1.15%
9	ل	3.44 %	19	ع	0.98%
10	ت	3.23%	20	ٿ	0.97%

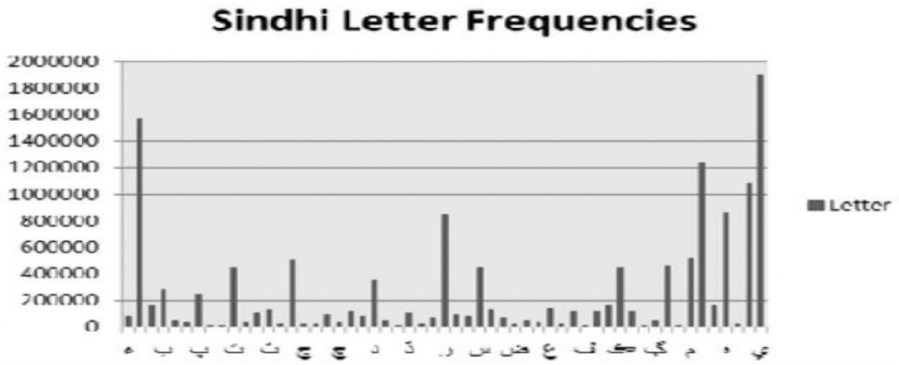

Figure 3. Sindhi Corpus Letter frequency distribution.

5.3. FREQUENCIES OF WORD

In this analysis of research paper we have examined and found 4.1 of millions words & 70,576 differ word forms. There were included most occuring words as container markers(like ۾ ،ڇ an ع) & incomplete / helping verbs like (آهيون and هو).

Table 4. Twenty frequent words in sindhi.

S.No.	Word	Percent	S.No.	Word	Percent
1	۾	2.42%	11	ڪري	0.68%
2	آهي	1.63%	12	ناس	0.69%
3	۽	2.17%	13	ان	0.66%
4	۾	1.78%	14	ناڪ	0.63%
5	يڻ	1.61%	15	ٽي	0.56%
6	ڪي	1.61%	16	نڻ	0.55%
7	وڃ	1.50%	17	ءِ	0.51%
8	ئي	1.05%	18	نه	0.50%
9	هڪ	0.82%	19	وهه	0.50%
10	نه	0.70%	20	ڪيو	0.45%

6. FUTURE WORK

The results are updated and achieved constantly for corpus. For specific POS tagging, n-gram based text classification for specific annotations the studies are in fast progress.

Table 5. Ten most frequent word bigrams.

S.No.	Word bigram	Percentage
1	۾ ويڃ	7.52
2	۾ يڻ	6.75
3	يڃ نه	2.66
4	وڏي ريظنيب	1.93
5	يڃ ڏنس	1.84
6	يڃ نا	1.72
7	۾ نهڙ	1.60
8	ويڃ نه	1.60
9	ويو ويڪ	1.44
10	يڻ ويو	1.21

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100-115

For advance enhancement and maturity of corpus the excess particular Sindhi computational linguistic studies are necessary and essential studies. Before tagging of POS the corpus the Sindhi tag set is to be required for designed. The areas to be extensively worked out which are quantitative improvement, proper annotation, qualitative and comprehensive statistical analysis.

7. CONCLUSION

In the fields of social science, applied science, computer science and other domains the research studies have brought major changes in different topics. It is a continuous process for the benefits for the development of society to make the things perfect. As for the research study is concerned the basic research study is done on analysis and development of Sindhi text corpus. For this purpose the Arabic-Persia script is used and simultaneously for the analysis of Sindhi text corpus the more research work is needed. For this purpose word 2 Vic, similarity analysis, sentiment analysis, topic modeling and cluster analysis are used. For future research the computational linguistics and NLP are contributed in Sindhi text corpora.

The Sindhi corpus construction project is very precious forward in language processing absence sources of Sindhi language. Despite of its magnitude and initial output of the corpus is present position will provide base for advance in studies of Sindhi language it is natural language process. For smart devices and cell phones the script frequencies which include bigram and trigram is providing base for compact keyboard design and intelligent text processing. For the correction of spelling and automatic sentence completion applications, word level unigram and bigram frequencies bring base. For enhanced language processing targets just like information retrieval and extraction and machine translation, semantic analysis, syntax analysis and morphological analysis, further enhancement in corpus will be very beneficial.

REFERENCES

Becker, D. & Riaz, K. (2002). A study in urdu corpus construction. In *Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12* (pp. 1-5). Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1118759.1118760

Decerbo, M., MacRostie, E. & Natarajan, P. (2004). *The BBN Byblos Pashto OCR system.*

In *Proceedings of the 1st ACM workshop on Hardcopy* document processing (pp. 29-32). ACM. doi: http://dx.doi.org/10.1145/1031442.1031447

Hakro, D. N., Ismaili, I. A., Talib, A. Z., Bhatti, Z. & Mojai, G. N. (2014). Issues and challenges in Sindhi OCR. *Sindh University Research Journal-SURJ (Science Series), 46*(2).

Hussain, S. (2008). Resources for Urdu language processing. In *Proceedings of the 6th workshop on Asian Language Resources.*

Hussain, S. & Durrani, N. (2008). *A study on collation of languages from developing Asia*. Center for Research in Urdu Language Processing, National University of Computer and Emerging Science, Lahore, PK.

Ijaz, M. & Hussain, S. (2007). Corpus based Urdu lexicon development. In the *Proceedings of Conference on Language Technology* (CLT07), University of Peshawar, Pakistan (Vol. 73).

Mahar, J. A. & Memon, G. Q. (2010). Rule based part of speech tagging of Sindhi language. In *2010 International Conference on Signal Acquisition and Processing* (pp. 101-106). IEEE.

Rahman, M. U. (2009). Sindhi morphology and noun inflections. In *Proceedings of the Conference on Language & Technology* (pp. 74-81).

Sindhi English Dictionary. Retrieved from http://www.crulp.org/sed/ (Accessed 2010).

DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100-115

Urdu, Nepali and English Parallel Corpus, CRULP. Rettrieved from http://crulp.org/software/ling_resources/Urdu Nepali EnglishP-arallelCorpus.htm (Accessed: 2010).

AUTHORS

Fida Hussain Khoso

Mr Khoso is perusing his Ph.D Computer Science, from Department of Computing, Faculty of Engineering, Science & Technology (FEST), Indus University Karachi Pakistan. He is working as a Lecturer at Dawood University of Engineering & Technology Karachi, Pakistan.

He has more than 06 research publications in national and international journals. His research area is Artificial Intelligence, NLP, Speech recognition system.

Mashooque Ahmed Memon

Mr. Memon working as a Lecturer in the Department of Computer Science and IT Benazir Bhutto Shaheed University Lyari Karachi

He has more than 10 research publications in national and international journals.

Haque Nawaz Lashari

Mr. Haque Nawaz Lashari is pursuing his PhD in Computer Science from Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Karachi, He received his MS degree from Mohammad Ali Jinnah University Karachi in Network and Telecommunication in 2010. He is working as Lecturer at Sindh Madressatul Islam University, Karachi. He has more than 24 research publications in national and international journals. His areas of research interests are wireless communication, network security, routing protocols, optimization algorithms and mobility management in mobile ad hoc networks

Prof. Dr. Engr. Sayed Hyder Abbas Musavi

Senior Member IEEE

Dr. Musavi earned his PhD Degree in 2011 in Telecommunication Engineering. He has 25 years of teaching and research experience. He is currently serving as Dean at Faculty of Engineering, Science & Technology Indus University, Karachi, Pakistan

Edición Especial Special Issue Mayo 2019
DOI: http://dx.doi.org/10.17993/3ctecno.2019.specialissue2.100-115