Vol. 2, No. 2 | July – December 2018

Developing Sindhi Text Corpus using XML Tags

Sayed Majid Ali Shah* Zeeshan Bhatti* Imdad Ali Ismaili*

Abstract

Sindhi language being one of the oldest languages of the world, has still very limited use in digital age due to lack of digital contents. The use of corpus for each language has been extremely important in facilitating the natural language processing of its script. This research work addresses the issue of building corpus for Sindhi Language using XML based Tagging. The tree based XML tag structure is designed to develop Sindhi Corpus that has two main nodes namely metadata and Sindhi Document which contains the main text. The Corpus developed contains a detailed metadata tags to represent Sindh language, documenting each relevant component of the corpus. The final corpa would be further used in various Natural language applications for Sindh language.

Keywords: *Corpus, Sindhi, Sindhi Corpus, Natural Language Processing, XML*

1. Introduction

Sindhi language is a widely spoken language based on Arabic script with similar cursive ligatures and written from right-to-left consisting of 52 characters [1] [2]. Sindhi language is considered as the second most popularly written and spoken language, after Urdu, in Pakistan. Even though Sindhi is an old language with vast amount of literature and written resources. However, there are very insufficient computational material and digital coprus available for Sindhi Language to create efficient NLP applications.

Natural Language Processing applications always require a huge collection of Corpus data for the language. A corpus is simply collection of large amount of structure and unstructured text for a language. The well-defined structural format is created to store and categorize the text in large datasets, allowing the computational processing and application development. This structured datasets facilitates the statistical analysis and grammatical validation of the script, along with other applications of NLP [3] [4].

Corpus are considered as one of the key prerequisites for and obligatory component for developing any Natural Language Processing applications such as, Spell checkers[2][5], Machine Learnings, Speech-to-Text, Text-to-Speech, OCR, Translation, Transliteration, etc. [6]. Due to this, there is huge need for developing a Sindhi language corpus which is also publicly available for everyone to use.

XML has always been a key technique for designing a structure for developing Corpa of various languages [7] [8] [9] [10]. XML is a very flexible language due to its tag-based structure, which allows the developer to easily extract the required and desired information from the structured XML document. Developing Sindhi corpus in XML would enable rule-based tagging's, and structured designing of Corpa, allowing an easier reusability of the corpus along with broader dissemination to various NLP applications.

Since Corpus is extremely essential for any language for NLP application, a huge amount of work from various aspects has been done to develop corpus for various different languages. Primarily, a project named EMILLE was developed consisting of multilingual corpora for South Asian languages [10]. Similarly, Urdu corpus was developed containing 18 million words by the Center for Research in Urdu Language Processing (CRULP) [11]. CRULP has also developed and released Online Urdu Dictionary (OUD) containing 120,000 records of Urdu corpus with 80,000 words dictionary words [9]. Whereas, Bank of English Corpus was developed to help the dictionaries [8]. On the other hand, Hindi Corpus was designed by IIT Bombay to facilitate the NLP development of the language [13] along with EMILLE corpus [12].

2. XML Structure for Sindhi Corpus

In NLP application development, XML tag-based structured format has been widely used to create structured

*Institute of Information and Communication Technology, University of Sindh, Jamshoro
Corresponding Email: lakiyarimajid@gmail.com

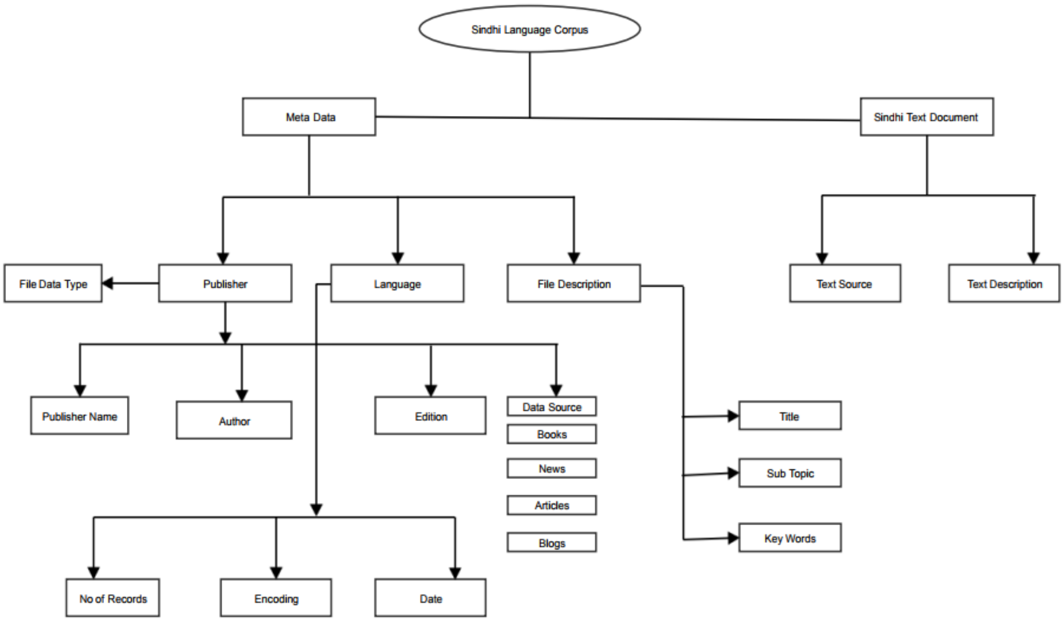


Figure 1: Proposed Model with Hypothesis relation

documents for processing and developing regional language applications [14]. For this purpose, XML has been used with custom tag to structure Sindhi text in a formalized Corpus for various NLP applications. The tags for Sindhi Corpus based on XML have been segregated into two main sections consisting of Metadata tags and Sindhi Text Document tags. Each is then further divided to contain more detailed information in its sub- tags.

A. Sindhi Corpus Structure

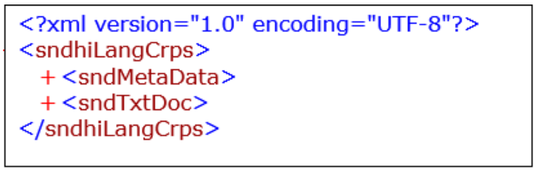
The XML based Sindhi Corpus structure has been divided into two main sections at the top-most level with 'Metadata' tag containing tags related to the original source information related to the actual text and document. The 'Sindhi Text Document' tag is second top-most tag containing the actual text from the source document. The full hierarchy of the XML tag structure for Sindhi Corpus is illustrated in Figure 1. Each Sindhi Corpa document will be stored with respect to this structure within XML tags.

B. Meta Data Header of Sindhi Corpus

Metadata is defined as the data about the data. Therefore, this main tag contains specific detail information related to the source document. This main section contains attributes such as "Title" of the document, "Sub Title" of Sindhi document, if any, "Topic" being discussed in the Sindhi document, "Sub Topic", "Book", "Author", "Edition", etc. The detailed sub-tag structure of the meta data section is shown in Figure 1. The 'Sindhi Text Document' tag contains the source raw text information which is extracted from various sources including websites, newpapers, books, articles, etc. This tag contains further two sub tags that describe the text description of the source text and the actual text file under "Text Description" and "Text Document" respectively.

3. Sindhi Corpus Representations

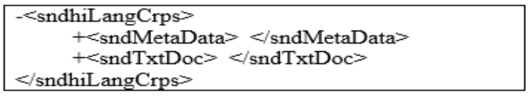
There are two main custom tags defined after <sndhiLangCrps> as <sndMetaData> </sndMetaData> and <sndTextDoc> </sndTextDoc>, and the operator (+) shows that both custom tags have also child tags as define operator (+) in Figure 2.



```
<?xml version="1.0" encoding="UTF-8"?>
<sndhiLangCrps>
  + <sndMetaData>
  + <sndTxtDoc>
</sndhiLangCrps>
```

Figure 2: Super tags of <sndhiLangCrps> SLC(Sindhi Language Corpus

Figure 3 shows the main Sindhi Language Corpus Tag that contains to top-most sub tags Sindh Meta Data nd Sindh Text Document tags.

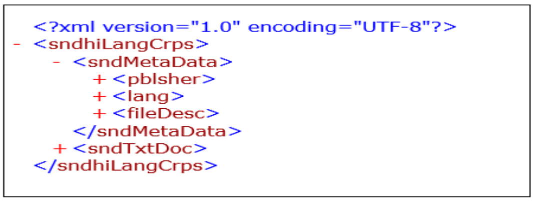


```
-<sndhiLangCrps>
        +<sndMetaData> </sndMetaData>
        +<sndTxtDoc> </sndTxtDoc>
</sndhiLangCrps>
```

Figure 3: Root and elements tags of sndhiLangCrps

A. Portion of SndMetaData

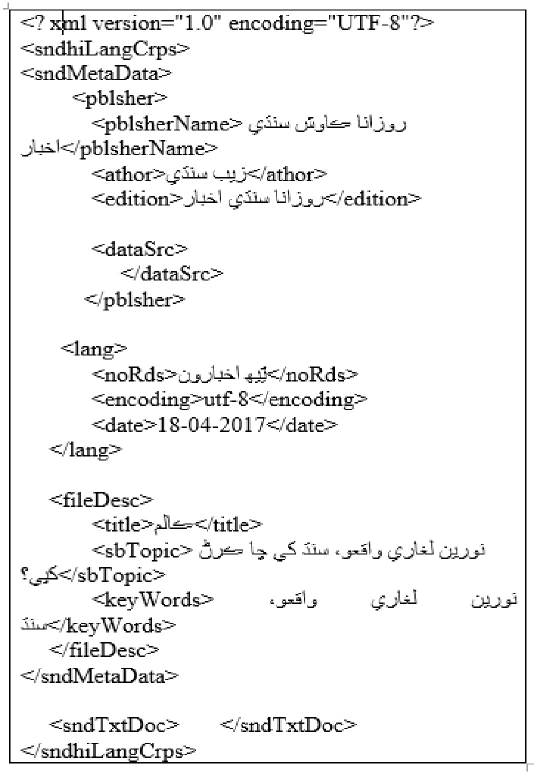
The tag of XML <sndMetaData> “Sindhi Meta Data” is the part of Sindhi corpus in data file which shows that the Sindhi data about its own data, it has child tags which also contains further information about the Sindhi document.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <sndhiLangCrps>
  - <sndMetaData>
      + <pblsher>
      + <lang>
      + <fileDesc>
    </sndMetaData>
  + <sndTxtDoc>
  </sndhiLangCrps>
```

Figure 4: Elements and child tags of SndMetaData at SndhiLangCrps

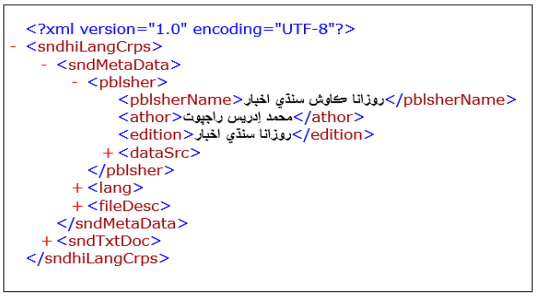
In this figure 5 the elements tags of <sndMetaData> </sndMetaData> has been defined as <pblsher> </pblsher>, <lang> </lang> and <fileDesc> </fileDesc> and they have also sub child as per operator defines.



```
<? xml version="1.0" encoding="UTF-8"?>
<sndhiLangCrps>
<sndMetaData>
      <pblsher>
         <pblsherName> روزانا ڪاوش سنڌي
اخبار</pblsherName>
         <athor>زيب سنڌي</athor>
         <edition>روزانا سنڌي اخبار</edition>

         <dataSrc>
             </dataSrc>
        </pblsher>

     <lang>
         <noRds>ٽيه اخبارون</noRds>
         <encoding>utf-8</encoding>
         <date>18-04-2017</date>
    </lang>

    <fileDesc>
         <title>ڪالم</title>
         <sbTopic> نورين لغاري واقعو، سنڌ کي ڇا ڪرڻ
کپي؟</sbTopic>
         <keyWords>     نورين     لغاري     واقعو،
سنڌ</keyWords>
    </fileDesc>
</sndMetaData>

    <sndTxtDoc>      </sndTxtDoc>
</sndhiLangCrps>
```

Figure 5: Sindhi XML Corpus with MetaData

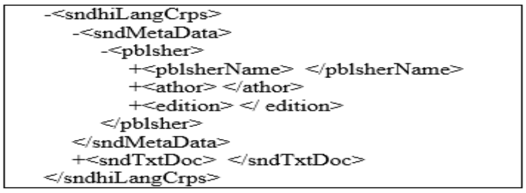
Figure 6 shows another example of Publisher tags data for Sindhi Document. In this figure, the child tags of <pblsher> </pblsher> has been defined as <pblsherName> </pblsherName>, <athor> </athor> and <edition> </edition> custom tags. Accurate data also filled in that custom tags for the building of SLC, while the other tags are here in silent mod they have discussed in other figure and the operator (-) shows that specific tag is displayed with own child and no more child is hide.



```
  <?xml version="1.0" encoding="UTF-8"?>
- <sndhiLangCrps>
  - <sndMetaData>
    - <pblsher>
        <pblsherName>روزانا ڪاوش سنڌي اخبار</pblsherName>
        <athor>محمد إدريس راجپوت</athor>
        <edition>روزانا سنڌي اخبار</edition>
      + <dataSrc>
      </pblsher>
    + <lang>
    + <fileDesc>
    </sndMetaData>
  + <sndTxtDoc>
  </sndhiLangCrps>
```

Figure 6: Data in publisher tag

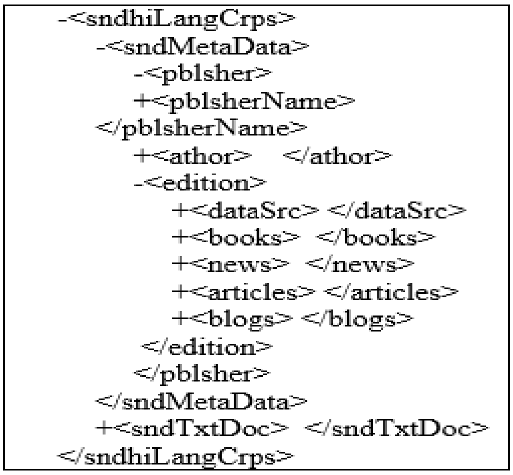
The XML tag <sndMetaData> has three elements tags named as <pblsher> Publisher, <lang> Language and <fileDesc> File Description.



```
-<sndhiLangCrps>
   -<sndMetaData>
      -<pblsher>
         +<pblsherName> </pblsherName>
         +<athor> </athor>
         +<edition> </ edition>
      </pblsher>
   </sndMetaData>
   +<sndTxtDoc> </sndTxtDoc>
</sndhiLangCrps>
```

Figure 7: Elements and child tags of pblsher in SndMetaData at SndhiLangCrps

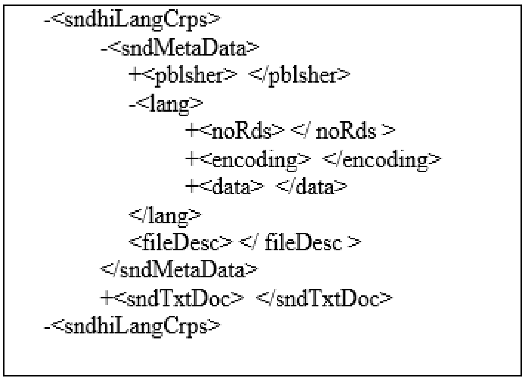
The publisher tag <pblsher> contains the information of publications, with describes its child elements as <pblsherName> “Publisher Name”, <athor> ”Author ” , <edition> “Edition”. The tag <pblsherName> shows the name of publisher, the tag <athor> tells the name of author while the tag <edition> describes the edition of publications.



```
-<sndhiLangCrps>
   -<sndMetaData>
      -<pblsher>
      +<pblsherName>
   </pblsherName>
      +<athor>   </athor>
      -<edition>
         +<dataSrc> </dataSrc>
         +<books>  </books>
         +<news>  </news>
         +<articles> </articles>
         +<blogs> </blogs>
       </edition>
      </pblsher>
   </sndMetaData>
   +<sndTxtDoc>  </sndTxtDoc>
</sndhiLangCrps>
```

Figure 8: Elements and child tags of pblsher with its child’s elements edition in SndMetaData at SndhiLangCrps

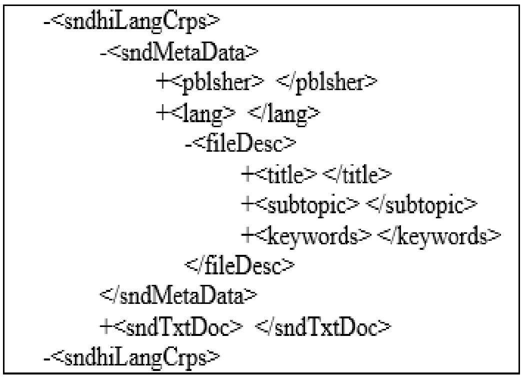
The tag <pblsher> publisher is the child tag of Sindhi Meta Data <sndMetaData> while the tag <edition> Edition is the child tag of <pblsher> and <edition> tag has its child as <dataSrc> “Data Source”, <books> “Books”, <news> “News”, <articles> “Articles”, <blogs> “Blogs”. These all are the sources of information which provide the complete data to edition and edition makes the complete to the publisher tag.



```
-<sndhiLangCrps>
      -<sndMetaData>
         +<pblsher> </pblsher>
         -<lang>
               +<noRds> </ noRds >
               +<encoding> </encoding>
               +<data> </data>
         </lang>
         <fileDesc> </ fileDesc >
      </sndMetaData>
      +<sndTxtDoc> </sndTxtDoc>
-<sndhiLangCrps>
```

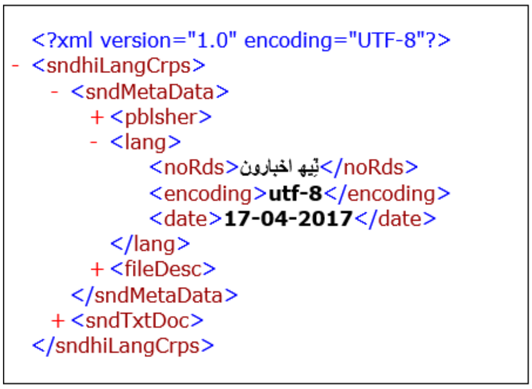
Figure 9: Elements and child tags of Lang in SndMetaData at SndhiLangCrps

The Language tag <lang> is the element tag of <sndMetaData> which has child tags like tag <noRds> “Number of Records”, <encoding> “Encoding”, <data> “Data”.



```
-<sndhiLangCrps>
      -<sndMetaData>
            +<pblsher> </pblsher>
            +<lang> </lang>
               -<fileDesc>
                     +<title> </title>
                     +<subtopic> </subtopic>
                     +<keywords> </keywords>
               </fileDesc>
      </sndMetaData>
      +<sndTxtDoc> </sndTxtDoc>
-<sndhiLangCrps>
```

Figure 10: Elements and child tags of fileDesc in SndMetaData at SndhiLangCrps

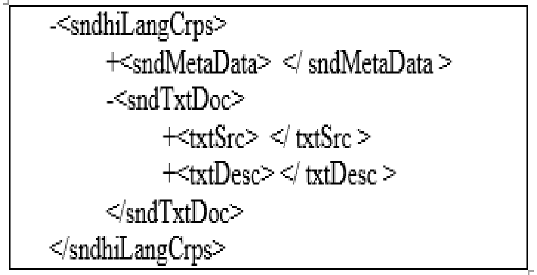


```
  <?xml version="1.0" encoding="UTF-8"?>
- <sndhiLangCrps>
  - <sndMetaData>
    + <pblsher>
    - <lang>
        <noRds>ٽيهه اخبارون</noRds>
        <encoding>utf-8</encoding>
        <date>17-04-2017</date>
      </lang>
    + <fileDesc>
    </sndMetaData>
  + <sndTxtDoc>
  </sndhiLangCrps>
```

Figure 11: child tags of <lang> </lang>

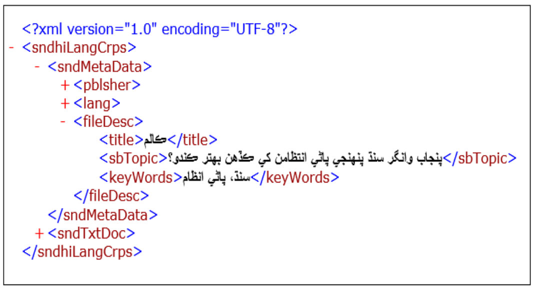
Figure 11 uses the child tags of <lang> </lang> tags as <noRds> </noRds>, <encoding> </encoding> and <date> </date>. That tags have filled by accurate data while other tags are here silent to show the role of that tags in corpus linguistics.

File Description <fileDesc> is the element tag of <sndMetaData> tag consists of child tags as <title> “Title”, <subtopic> “Sub Topic”, <keywords> “Key Words”.



```
-<sndhiLangCrps>
      +<sndMetaData> </ sndMetaData >
      -<sndTxtDoc>
            +<txtSrc> </ txtSrc >
            +<txtDesc> </ txtDesc >
      </sndTxtDoc>
</sndhiLangCrps>
```

Figure 12: Elements and child tags of SndTxtDoc at SndhiLangCrps



```
  <?xml version="1.0" encoding="UTF-8"?>
- <sndhiLangCrps>
  - <sndMetaData>
    + <pblsher>
    + <lang>
    - <fileDesc>
        <title>ڪالم</title>
        <sbTopic>پنجاب وانگر سنڌ پنهنجي پاڻي انتظامن کي ڪڏهن بهتر ڪندو؟</sbTopic>
        <keyWords>سنڌ، پاڻي انظام</keyWords>
      </fileDesc>
    </sndMetaData>
  + <sndTxtDoc>
  </sndhiLangCrps>
```

Figure 13: Element tags of <fileDesc> </fileDesc>

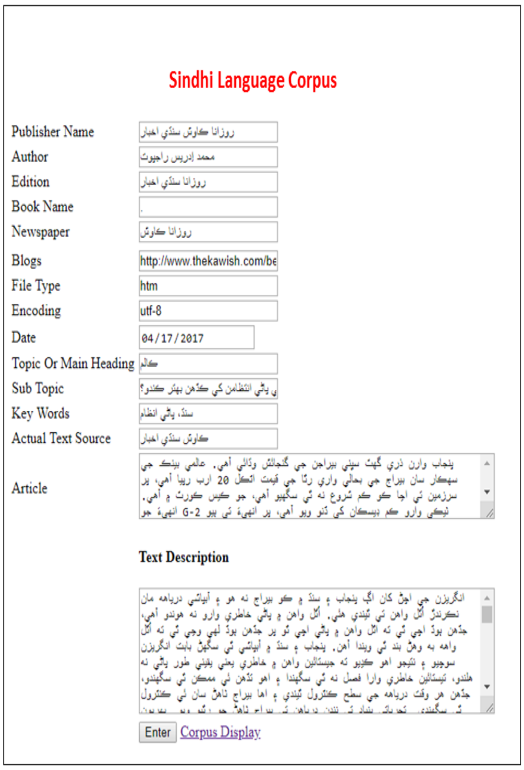
Figure 13 shows the sub tags of custom tag of <fileDesc> </fileDesc> as <title> </title>, <sbTopic> </sbTopic> and <keyWords> </keyWords>. All tags have assigned their own data.

B. Portion of Sindhi Text Document

The tag of XML <sndTxtDoc> “Sindhi Text Document” is the part of Sindhi corpus in data file, it has also child tags as <txtSrc> “Text Source”, <txtDesc> “Text Description”.

4. Sample Sindhi Corpus Document

The final Sindhi documents are initially created manually by extracting information form articles and saving them in XML tags as discussed [15]. A GUI form was designed that allowed the creating of XML document for Sindhi text as shown in Figure 14. Each entry was saved as an XML file as per rules and patterns discussed above.



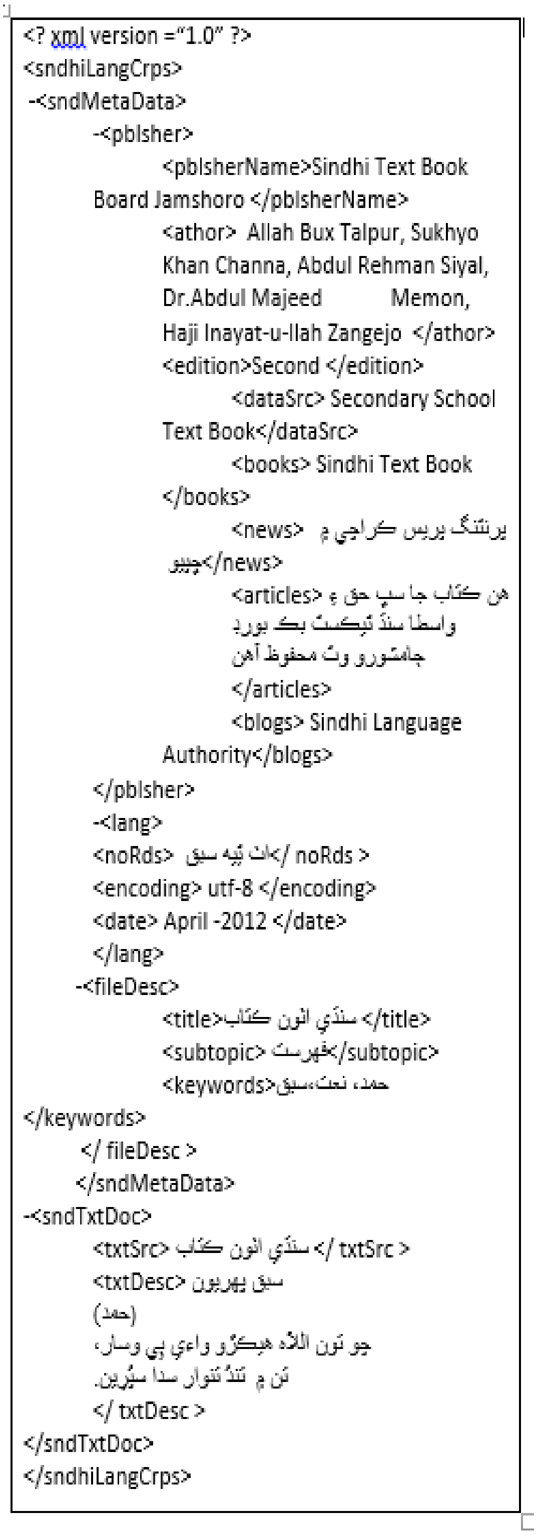
Sindhi Language Corpus

Publisher Name	روزانا ڪاوش سنڌي اخبار
Author	محمد إدريس راجپوت
Edition	روزانا سنڌي اخبار
Book Name	.
Newspaper	روزانا ڪاوش
Blogs	http://www.thekawish.com/be
File Type	htm
Encoding	utf-8
Date	04/17/2017
Topic Or Main Heading	ڪالم
Sub Topic	ي پاڻي انتظامن کي ڪڏهن بهتر ڪندو؟
Key Words	سنڌ، پاڻي انظام
Actual Text Source	ڪاوش سنڌي اخبار
Article	پنجاب وارن ذري گهٽ سيني بيراجن جي گنجائش وڌائي آهي. عالمي بينڪ جي سيهڪار سان بيراج جي بحالي واري رٿا جي قيمت اٽڪل 20 ارب رپيا آهي، پر سرزمين تي اڃا ڪو ڪم شروع نه ٿي سگهيو آهي، جو ڪيس ڪورٽ ۾ آهي. ٺيڪي وارو ڪم ڊيسڪان کي ڏنو ويو آهي، پر انهيءَ تي پيو G-2 انهيءَ جو

Text Description

انگريزن جي اچڻ کان اڳ پنجاب ۽ سنڌ ۾ ڪو بيراج نه هو ۽ آبپاشي درياهه مان
نڪرندڙ اٽل واهن تي ٿيندي هئي. اٽل واهن ۾ پاڻي خاطري وارو نه هوندو آهي،
جڏهن ٻوڏ اچي ٿي ته اٽل واهن ۾ پاڻي اچي ٿو پر جڏهن ٻوڏ لهي وڃي ٿي ته اٽل
واهه به وهڻ بند ٿي ويندا آهن. پنجاب ۽ سنڌ ۾ آبپاشي ٿي سگهڻ بابت انگريزن
سوچيو ۽ نتيجو اهو ڪڍيو ته جيستائين واهن ۾ خاطري يعني يقيني طور پاڻي نه
هلندو، تيستائين خاطري وارا فصل نه ٿي سگهندا ۽ اهو تڏهن ئي ممڪن ٿي سگهندو،
جڏهن هر وقت درياهه جي سطح ڪنٽرول ٿيندي ۽ اها بيراج ٺاهڻ سان ئي ڪنٽرول

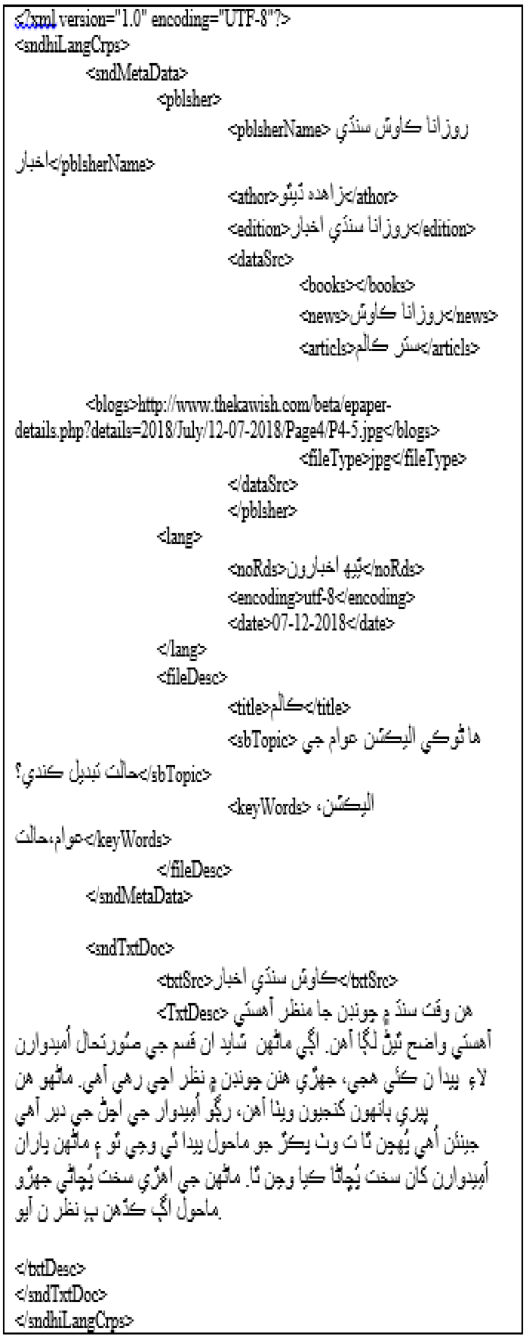
Figure 14: GUI form for creating XML document of Sindhi corpus



```
<? xml version ="1.0" ?>
<sndhiLangCrps>
 -<sndMetaData>
          -<pblsher>
                    <pblsherName>Sindhi Text Book
          Board Jamshoro </pblsherName>
                    <athor>  Allah Bux Talpur, Sukhyo
                    Khan Channa, Abdul Rehman Siyal,
                    Dr.Abdul Majeed          Memon,
                    Haji Inayat-u-llah Zangejo  </athor>
                    <edition>Second </edition>
                              <dataSrc> Secondary School
                    Text Book</dataSrc>
                              <books> Sindhi Text Book
                    </books>
                              <news>  پرنٽنگ پريس ڪراچي ۾
                     ڇپيو</news>
                              <articles> هن ڪتاب جا سڀ حق ۽
                              واسطا سنڌ ٽيڪسٽ بڪ بورڊ
                              ڄامشورو وٽ محفوظ آهن
                              </articles>
                              <blogs> Sindhi Language
                    Authority</blogs>
          </pblsher>
          -<lang>
          <noRds>  ڪاٺ ٽيه سبق</ noRds >
          <encoding> utf-8 </encoding>
          <date> April -2012 </date>
          </lang>
        -<fileDesc>
                    <title>سنڌي اٺون ڪتاب</title>
                    <subtopic> فهرست</subtopic>
                    <keywords>حمد، نعت،سبق
</keywords>
         </ fileDesc >
        </sndMetaData>
-<sndTxtDoc>
          <txtSrc> سنڌي اٺون ڪتاب </ txtSrc >
          <txtDesc> سبق پهريون
          (حمد)
          جو تون الله هيڪڙو واءِ ٻي وسار،
          تن ۾ تندُ تنوار سدا سيڙين.
          </ txtDesc >
</sndTxtDoc>
</sndhiLangCrps>
```

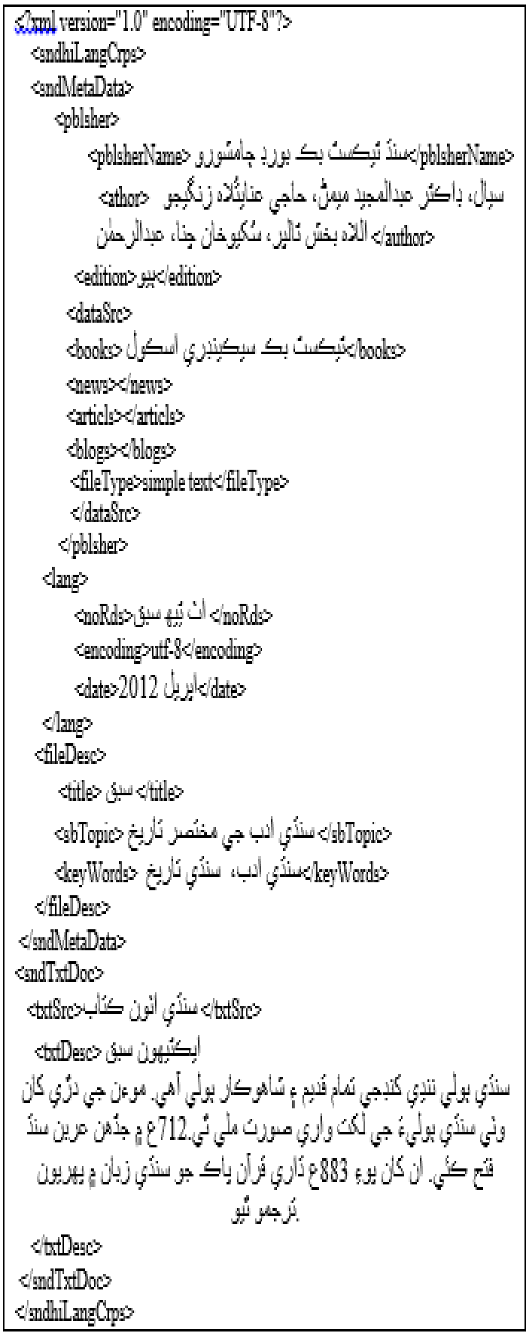
Figure 15: Sample Sindhi XML document

The final version of each XML document of Sindhi corpus contain all the relevant information that could easily be read and processed for any NLP task as shown in Figure 15 to Figure 18.



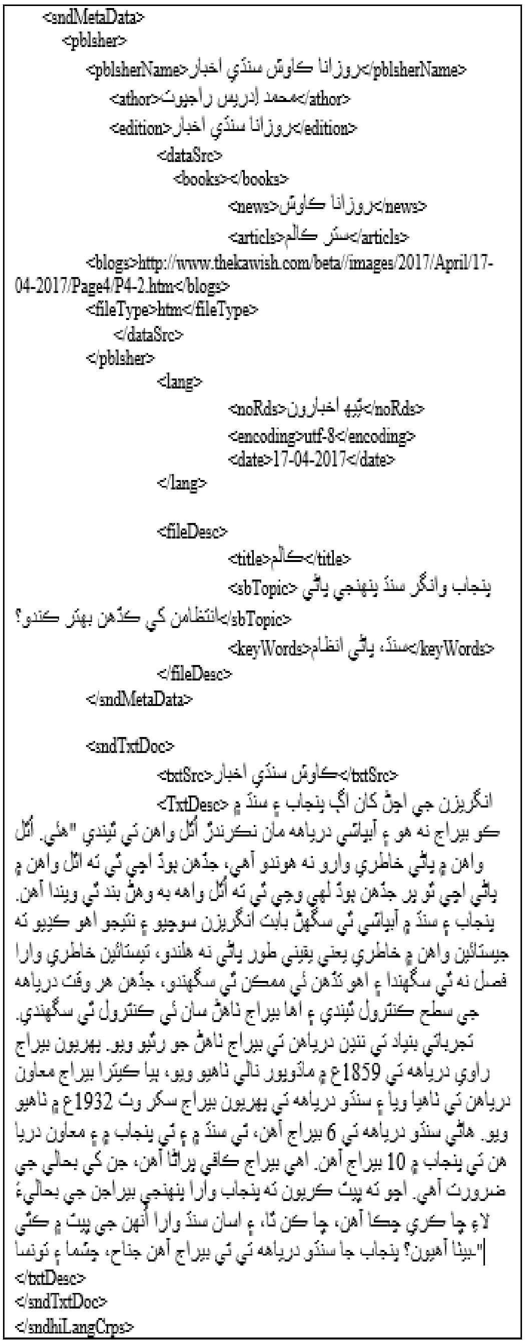
```
<?xml version="1.0" encoding="UTF-8"?>
<sndhiLangCrps>
        <sndMetaData>
                <pblsher>
                        <pblsherName> روزانا ڪاوش سنڌي
اخبار</pblsherName>
                        <athor>زاهده ڏيٿو</athor>
                        <edition>روزانا سنڌي اخبار</edition>
                        <dataSrc>
                                <books></books>
                                <news>روزانا ڪاوش</news>
                                <articls>سنڌ ڪالم</articls>

        <blogs>http://www.thekawish.com/beta/epaper-
details.php?details=2018/July/12-07-2018/Page4/P4-5.jpg</blogs>
                                <fileType>jpg</fileType>
                        </dataSrc>
                        </pblsher>
                <lang>
                        <noRds>ٽيهه اخبارون</noRds>
                        <encoding>utf-8</encoding>
                        <date>07-12-2018</date>
                </lang>
                <fileDesc>
                        <title>ڪالم</title>
                        <sbTopic> ها ٿوڪي اليڪشن عوام جي
حالت تبديل ڪندي؟</sbTopic>
                        <keyWords> اليڪشن،
عوام،حالت</keyWords>
                </fileDesc>
        </sndMetaData>

        <sndTxtDoc>
                <txtSrc>ڪاوش سنڌي اخبار</txtSrc>
                <TxtDesc> هن وقت سنڌ ۾ چونڊن جا منظر آهستي
آهستي واضح ٿيڻ لڳا آهن. اڳي ماڻهن شايد ان قسم جي صورتحال اُميدوارن
لاءِ پيدا نه ڪئي هجي، جهڙي هن چونڊن ۾ نظر اچي رهي آهي. ماڻهو هن
پيري پانهون ڪنجيون ويٺا آهن، رڳو اُميدوار جي اچڻ جي دير آهي
جيئن اهي پُهچن ٿا ته وٺ پڪڙ جو ماحول پيدا ٿي وڃي ٿو ۽ ماڻهن پاران
اُميدوارن کان سخت پُڇاڻا ڪيا وڃن ٿا. ماڻهن جي اهڙي سخت پُڇاڻي جهڙو
ماحول اڳ ڪڏهن به نظر نه آيو.

</txtDesc>
</sndTxtDoc>
</sndhiLangCrps>
```

Figure 16: Sindhi XML Document 2



```
<?xml version="1.0" encoding="UTF-8"?>
  <sndhiLangCrps>
  <sndMetaData>
     <pblsher>
         <pblsherName> سنڌ ٽيڪسٽ بڪ بورڊ ڄامشورو</pblsherName>
           <athor>  سيال، ڊاڪٽر عبدالمجيد ميمڻ، حاجي عنايتُلاه زنگيجو
           الاه بخش ٿالير، سڪيوخان جتا، عبدالرحمان </author>
        <edition>ڇيو</edition>
       <dataSrc>
       <books> ٽيڪسٽ بڪ سيڪنڊري اسڪول</books>
       <news></news>
       <articls></articls>
       <blogs></blogs>
        <fileType>simple text</fileType>
        </dataSrc>
      </pblsher>
    <lang>
        <noRds>اٺ ٽيهه سبق </noRds>
        <encoding>utf-8</encoding>
        <date>2012 اپريل</date>
    </lang>
   <fileDesc>
      <title> سبق </title>
     <sbTopic> سنڌي ادب جي مختصر تاريخ </sbTopic>
     <keyWords>  سنڌي ادب، سنڌي تاريخ</keyWords>
   </fileDesc>
 </sndMetaData>
<sndTxtDoc>
  <txtSrc>سنڌي اٺون ڪتاب </txtSrc>
   <txtDesc> ايڪٽيهون سبق
 سنڌي ٻولي ننڍي کنڊجي تمام قديم ۽ شاهوڪار ٻولي آهي. موهن جي دڙي کان
 وٺي سنڌي ٻوليءَ جي لکت واري صورت ملي ٿي.712ع ۾ جڏهن عربن سنڌ
  فتح ڪئي. ان کان پوءِ 883ع ڌاري قرآن پاڪ جو سنڌي زبان ۾ پهريون
                        ترجمو ٿيو.
   </txtDesc>
 </sndTxtDoc>
</sndhiLangCrps>
```

Figure 17: Sample Sindhi XML Document 3



```
    <sndMetaData>
        <pblsher>
            <pblsherName>روزانا ڪاوش سنڌي اخبار</pblsherName>
                <athor>محمد ادريس راجپوت</athor>
                <edition>روزانا سنڌي اخبار</edition>
                        <dataSrc>
                            <books></books>
                                    <news>روزانا ڪاوش</news>
                                    <articls>سنڌ ڪالم</articls>
            <blogs>http://www.thekawish.com/beta//images/2017/April/17-
04-2017/Page4/P4-2.htm</blogs>
            <fileType>htm</fileType>
                 </dataSrc>
            </pblsher>
                        <lang>
                                    <noRds>ٽيهه اخبارون</noRds>
                                    <encoding>utf-8</encoding>
                                    <date>17-04-2017</date>
                        </lang>

                        <fileDesc>
                                    <title>ڪالم</title>
                                    <sbTopic> پنجاب وانگر سنڌ پنهنجي پاڻي
انتظامن کي ڪڏهن بهتر ڪندو؟</sbTopic>
                                    <keyWords>سنڌ، پاڻي انتظام</keyWords>
                        </fileDesc>
            </sndMetaData>

            <sndTxtDoc>
                        <txtSrc>ڪاوش سنڌي اخبار</txtSrc>
                        <TxtDesc> انگريزن جي اچڻ کان اڳ پنجاب ۽ سنڌ ۾
ڪو بيراج نه هو ۽ آبپاشي درياهه مان نڪرندڙ اُٿل واهن تي ٻڌل "هئي. اُٿل
واهن ۾ پاڻي خاطري وارو نه هوندو آهي، جڏهن ٻوڏ اچي ٿي ته اٿل واهن ۾
پاڻي اچي ٿو پر جڏهن ٻوڏ لهي وڃي ٿي ته اُٿل واهه به وهڻ بند ٿي ويندا آهن.
پنجاب ۽ سنڌ ۾ آبپاشي ٿي سگهڻ بابت انگريزن سوچيو ۽ نتيجو اهو ڪڍيو ته
جيستائين واهن ۾ خاطري يعني يقيني طور پاڻي نه هلندو، تيستائين خاطري وارا
فصل نه ٿي سگهندا ۽ اهو تڏهن ئي ممڪن ٿي سگهندو، جڏهن هر وقت درياهه
جي سطح ڪنٽرول ٿيندي ۽ اها بيراج ٺاهڻ سان ئي ڪنٽرول ٿي سگهندي.
تجرباتي بنياد تي ننڍن درياهن تي بيراج ٺاهڻ جو رٿيو ويو. پهريون بيراج
راوي درياهه تي 1859ع ۾ ماڌوپور نالي ٺاهيو ويو، پيا ڪيترا بيراج معاون
درياهن تي ٺاهيا ويا ۽ سنڌو درياهه تي پهريون بيراج سکر وٽ 1932ع ۾ ٺاهيو
ويو. هاڻي سنڌو درياهه تي 6 بيراج آهن، ٽي سنڌ ۾ ۽ ٽي پنجاب ۾ ۽ معاون دريا
هن تي پنجاب ۾ 10 بيراج آهن. اهي بيراج ڪافي پراڻا آهن، جن کي بحالي جي
ضرورت آهي. اڄ نه پيٽ ڪريون ته پنجاب وارا پنهنجي بيراجن جي بحاليءَ
لاءِ ڇا ڪري چڪا آهن، ڇا ڪن ٿا، ۽ اسان سنڌ وارا اُنهن جي پيٽ ۾ ڪٿي
"بيٺا آهيون؟ پنجاب جا سنڌو درياهه تي ٽي بيراج آهن جناح، چشما ۽ ٽونسا|
</txtDesc>
</sndTxtDoc>
</sndhiLangCrps>
```

Figure 18: Sample Sindhi XML Document 4

5. Conclusion and Future Work

The use of corpus in Natural Language Processing is extremely essential and important. The Sindhi Language Corpus is designed using XML tags to facilitate the processing of Sindhi text for various NLP tasks. XML tags have been designed to provide maximum data facilitation and a long term usability of Sindhi Corpus. The tab structure is segregated into two main sections containing metadata and main source full document. The metadata is crucial part of any document, and so Sindhi corpus metadata also contains many sub tags to cater for all possible information of any document. The use of XML for Sindhi corpus has been very fruitful and has provided a platform to work on more processing of Sindhi Text.

6. Acknowledgment

An earlier version of this paper was presented at the International Conference on Computing, Mathematics and Engineering Technologies (iCoMET 2018) and was published in its Proceedings available at IEEE Explorer, available at (URL: https://ieeexplore.ieee.org/iel7/8337998/8346308/08346381.pdf) This research work was carried out in Multimedia Animation and Graphics (MAGic) Research Group at Institute of Information and Communication Technology, University of Sindh, Jamshoro.

References

[1] Ismaili, I. A., Bhatti, Z., & Shah, A. A. (2014). Design & Development of the Graphical User Interface for Sindhi Language. arXiv preprint arXiv:1401.1486.

[2] Bhatti, Z., Waqas, A., Ismaili, I. A., Hakro, D. N., & Soomro, W. J. (2014). Phonetic based soundex & shapeex algorithm for sindhi spell checker system. arXiv preprint arXiv:1405.3033.

[3] Rahman, M. U. (2010). Towards Sindhi corpus construction. In Conference on Language and Technology, Lahore, Pakistan.

[4] Ko, W. K., & Phyo, T. Z. (2008, January). Selection of XML tag set for Myanmar National Corpus. In IJCNLP (pp. 33-40).

[5] Bhatti, Z., Ali Ismaili, I., Nawaz Hakro, D., & Javid Soomro, W. (2015). Phonetic-based sindhi spellchecker system using a hybrid model. Digital Scholarship in the Humanities, 31(2), 264-282.

[6] 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).

[7] Mahar, J. A., & Memon, G. Q. (2010, February). Rule based part of speech tagging of sindhi language. In Signal Acquisition and Processing, 2010. ICSAP'10. International Conference on (pp. 101-106). IEEE.

[8] Sinclair J. (1992), Introduction. BBC English Dictionary, London: Harper Collins. Tony M. and Wilson A. (2001), Corpus Linguistics (Second Edition), Edinburgh University Press.

[9] Rahman, S. (2005). Lexical Content and Design Case Study. Presented at From Localization to Language Processing, Second Regional Training of PAN Localization Project. Online presentation version: http://panl10n.net/Presentations/Cambodia/Shafiq/LexicalContent&Design.pdf.

[10] McEnery, A., Baker, J., Gaizauskas, R. & Cunningham, H. (2000). EMILLE: towards a corpus of South Asian languages, British Computing Society Machine Translation Specialist Group, London, UK.

[11] Ijaz, M. and Hussain, S. 2007. Corpus Based Urdu Lexicon Development. The Proceedings of Conference on Language Technology (CLT07), University of Peshawar, Pakistan.

[12] Hardie, A., Baker, P., McEnery, T., & Jayaram, B. D. (2006). Corpus-building for South Asian languages. TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS, 175, 211.

[13] Bojar, O., Diatka, V., Rychlý, P., Straňák, P., Suchomel, V., Tamchyna, A., & Zeman, D. (2014, May). HindEnCorp-Hindi-English and Hindi-only Corpus for Machine Translation. In LREC (pp. 3550-3555).

[14] Kim, J. D., Ohta, T., Tateisi, Y., Mima, H., & Tsujii, J. I. (2001). XML-based linguistic annotation of corpus. In Proc. of the First NLP and XML Workshop.

[15] Shah, S. M. A., Bhatti, Z., Ismaili, I. A., & Waqas, A. Designing XML tag based Sindhi Language Corpus. International Conference on Computing, Mathematics and Engineering Technologies – iCoMET 2018. IEEEXplore