A Study in Urdu Corpus Construction

Dara Becker
Graduate Program in Software Engineering
University of St. Thomas
St. Paul, MN, 55105, U.S.A.
dmbecker@stthomas.edu

Kashif Riaz
Department of Computer Science
University of Minnesota-Twin Cities
Minneapolis, MN, 55455, U.S.A.
riaz@cs.umn.edu

Abstract

We are interested in contributing a small, publicly available Urdu corpus of written text to the natural language processing community. The Urdu text is stored in the Unicode character set, in its native Arabic script, and marked up according to the Corpus Encoding Standard (CES) XML Document Type Definition (DTD). All the tags and metadata are in English. To date, the corpus is made entirely of data from British Broadcasting Company's (BBC) Urdu Web site, although we plan to add data from other Urdu newspapers. Upon completion, the corpus will consist mostly of raw Urdu text marked up only to the paragraph level so it can be used as input for natural language processing (NLP) tasks. In addition, it will be hand-tagged for parts of speech so the data can be used to train and test NLP tools.

Introduction

We are interested in contributing a small, publicly available Urdu corpus of written text to the natural language processing community. In pursuit of natural language processing research in Urdu, we could not find a publicly available Urdu corpus with which to work, so we had to start our own to train and test machine learning algorithms.

The language engineering community seems anxious to move forward fast in research of South Asian languages, but cannot because corpora of South Asian languages are not ample. "There is a dearth of work on Indic languages. The need to focus on Indic languages was further strengthened by our major review (with over 80 research centres world wide responding) of the needs of the [language engineering] community. Indic languages are the ones that most researchers want to work with but cannot because lack of corpus resources" [1].

1 Urdu corpus

Our corpus is currently made up of newspaper articles and columns from the Urdu Internet site of the British Broadcasting Company (BBC Urdu). News story data is easy to gather because it is readily available on the Internet and already in electronic form, although Web sites in Urdu tend to be published in graphics (a point we will return to later).

It is important for the users of corpus data to know from where the data came. Software trained on a written text corpus will perform poorly on spoken data and vice versa.

Something to keep in mind when using this Urdu data is that vocabulary and the stylistics of news stories in Urdu are very different than in everyday speech. For example, in Urdu news stories, "militants" are described as "people who like violence" تشدد پسند عناصر . Such a phrase is hardly ever used in everyday speech. Headings of news stories have different stylistics. For example, a common way to associate a statement to the person who made the statement is to write the statement followed by a colon or dash and then the person's name. This trend has been observed in Urdu news stories published in Pakistan, India, the UK, and in the United States.

The first version of the Urdu corpus to be published will be relatively small (20,000–50,000 words), but we will regularly be adding to the corpus as time passes. It will publicly available at http://personal1.stthomas.edu/dmbecker/.

All the Urdu documents will appear in a minimally tagged format (i.e., only paragraph tags) and, in addition, will be hand-tagged for parts of speech.

2 XML

The natural choice these days for storing a corpus is in an XML format. An XML format provides needed standardization so that a user who is unfamiliar with the corpus data, but familiar with a given XML DTD, can interface with the corpus fairly efficiently. At its best, software that has been previously designed to handle a corpus marked up in a given XML structure can handle a new corpus marked up in the same structure. This is advantageous because someone does not have to comb through the new corpus trying to understand its design in order to redesign the software that interfaces with the corpus. The designer of a corpus is always familiar with his/her own design, so one advantage of using an XML language to mark up a corpus is to make the corpus readily available to other researchers.

We chose the Corpus Encoding Standard (CES) XML DTD to mark up our corpus [2]. The main enclosing tag in this DTD is <cesCorpus> which is broken into main parts, <cesHeader> and <cesDoc>.

The header <cesHeader> contains meta information about the corpus data such as, date created, creator's name and contact information, description of the source, categories of the content, the writing system of the language being stored, how hyphenation in the source text is handled, and much more information (Figure 1).

The document tag <cesDoc> is where the actual text of the language of interest is stored. Each document is itself marked up with metadata specific to each document, like topic and source information for every separate document in the corpus.

The language data inside the <cesDoc> tags can be marked up simply with a paragraph tag <p> (Figure 2) or they can be more elaborately marked up with tags of semantic value (e.g., date, number, measure, name, term, time, foreign word) and formatting value (e.g., figure, table, p, sp, div, caption) (Figure 3). Tags that indicate formatting features such as 'caption' are important because they can be used, for example, to automatically determine the topic of a story.

The actual implementation of tagging Urdu script at a detailed level presents a display problem for our XML editor of choice, XML Spy. Upon looking at Figure 3, which is an excerpt from XML Spy, one may think that the word order of the paragraph is out of order. At the display level, the word order is out of order—it is barely human-readable, but at the storage level, the text is perfectly tagged and will process correctly. In Figure 4, we show, in a human-readable format, the order in which the Urdu text and English tags are stored. If an XML editor were optimized to display a right-to-left language with left-to-right tags, this is how we imagine the text would look. More importantly though, this is the order in which XML Spy currently stores the Urdu corpus.

We began the corpus building process by storing Urdu documents at the paragraph level with no other tags peppering the data. However, we intend to hand tag the data for parts of speech so the data can be used to train and test natural language processing algorithms.

```
<cesHeader type="corpus" creator="Dara Becker" version="1.0" status="update"
    date.created="2/2/02" date.updated="4/17/02">
    <fileDesc>
        <titleStmt><h.title>Urdu Corpus</h.title></titleStmt>
        <editionStmt version="1.0a"/>
        <publicationStmt>
            <distributor>Dara Becker</distributor>
            <telephone></telephone>
            <eAddress type="email">dmbecker@stthomas.edu</eAddress>
            <eAddress type="www">http://personal1.stthomas.edu/DMBECKER/</eAddress>
            <availability status="free"/>
        </publicationStmt>
    </fileDesc>
</cesHeader>
```

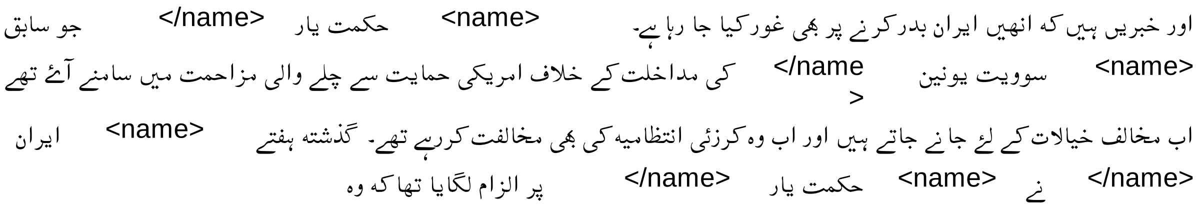
Figure 1: An excerpt from the corpus header
(It is not well-formed because we deleted some required tags.)

```
<cesDoc version="1.0">
     <cesHeader type="text" creator="Dara Becker" version="1.0" status="new"
          date.created="2/18/02" date.updated="" lang="ur">
     </cesHeader>
     <text>
          <body>
               <p>
                    <title>غور پر کرنے بدر ایران کو یار حکمت<title>
               </p>
               <p>
امریکی خلاف کے مداخلت کی یونین سوویت سابق جو یار حکمت ہے۔ رہا جا کیا غور بھی پر کرنے بدر ایران انھیں کہ ہیں خبریں اور
بھی کی انتظامیہ کرزئی وہ اب اور ہیں جاتے جانے لئے کے خیالات مخالف اب تھے آئے سامنے میں مزاحمت والی چلے سے حمایت
کاروائیاں خلاف کے انتظامیہ افغان کو سرزمین کی ایران وہ کہ تھا لگایا الزام پر یار حکمت نے ایران ہفتے گذشتہ تھے۔ کررہے مخالفت
وہ تھا رہا کر فراہم حمایت جو کو دھڑوں مزاحم خلاف کے طالبان وہ کہ ہے کہنا کا ایران کہ جب ہیں کررہے استعمال لئے کے کرنے
کے امریکہ اقدام خلاف کے یار حکمت نے ایران کہ ہے خیال کا ذرائع بعض تاہم ہے۔ گئی دی کر بند بعد کے ہونے ختم کنٹرول کا طالبان
ہیں۔ کیے بعد کے اعتراضات
               </p>
          </body>
     </text>
</cesDoc>
```

Figure 2: An excerpt from a corpus document
(It is not well-formed because we deleted some required tags.)

```
<text>
          <body>
                    <p>
                              <title>غور پر کرنے بدر ایران کو یار حکمت</title>
                    </p>
                    <p>یار حکمت<name> ہے۔ رہا جا کیا غور بھی پر کرنے بدر ایران انھیں کہ ہیں خبریں اور
والی چلے سے حمایت امریکی خلاف کے مداخلت کی </name>یونین سوویت <name>سابق جو </name>□ر
مخالفت بھی کی انتظامیہ کرزئی وہ اب اور ہیں جاتے جانے لئے کے خیالات مخالف اب تھے آئے سامنے میں مزاحمت
<name>وہ کہ تھا لگایا الزام پر </name>یار حکمت<name> نے </name>ایران<name> ہفتے گذشتہ تھے۔ کررہے
<name>کہ جب ہیں کررہے استعمال لئے کے کرنے کاروائیاں خلاف کے انتظامیہ افغان کو سرزمین کی </name>ایران
تھا رہا کر فراہم حمایت جو کو دھڑوں مزاحم خلاف کے </name>طالبان<name> وہ کہ ہے کہنا کا </name>ایران
کہ ہے خیال کا ذرائع بعض تاہم ہے۔ گئی دی کر بند بعد کے ہونے ختم کنٹرول کا </name>طالبان<name> وہ
اعتراضات کے </name>امریکہ<name> اقدام خلاف کے </name>یار حکمت <name>نے </name>ایران<name>
ہیں۔ کیے بعد کے</p>
          </body>
</text>
```

Figure 3: An illustration of how detailed tagging rearranges the display of the text in XML Spy



اور خبریں ہیں کہ انھیں ایران بدر کرنے پر بھی غور کیا جا رہا ہے۔ <name> حکمت یار </name> جو سابق
<name> سوویت یونین </name> کی مداخلت کے خلاف امریکی حمایت سے چلے والی مزاحمت میں سامنے آئے تھے
اب مخالف خیالات کے لئے جانے جاتے ہیں اور اب وہ کرزئی انتظامیہ کی بھی مخالفت کررہے تھے۔ گذشتہ ہفتے <name> ایران
</name> نے <name> حکمت یار </name> پر الزام لگایا تھا کہ وہ

Figure 4: A human-readable rendition of what tagged Urdu would look like in an XML editor optimized to display a right-to-left language with left-to-right tags

3 Unicode

Another natural choice for storing data is to use the Unicode character set. The Unicode character set is another needed standard that we take advantage of in order to make our corpus data readily available to other researchers.

The only reason for choosing to initially store text from BBC Urdu, and not other news agencies, is that the BBC publishes in the Unicode character set. Other news sites that publish in Urdu have gotten in the habit of publishing in graphics, presumably to avoid the hassles of arranging compatible fonts and character sets in the publishing software, systems, and client browsers. We think too it could be that Urdu publishers prefer Nastaliq-style font. There are probably a host of wonderful Nastaliq-style fonts available that work on legacy character sets, and, perhaps, publishers prefer to keep using these fonts.

The choice to publish in graphics though makes it difficult for data harvesters to snag data from the Web. If one really wants the data that are published in graphic form, one has to rekey the text, scan it using optical character recognition technology, or contact the publisher for electronic copies of text, in which case one needs to be able to handle or convert from the character set in which the text was originally typed. In a previous project, we developed an application that can convert between 120 legacy character sets and can be customized to convert any other font or character set, so we should have minimal obstacles when it comes time to harvest non-Unicode data.

Storing Urdu data in the Unicode character set eliminates some problems—however, we have found other problems related to different approaches to mapping Unicode-based fonts to the Arabic subset of Unicode.

Unicode-based fonts seem to have been optimized for Arabic display, not for Urdu, so we have experienced difficulty displaying various forms of *heh*, *noon ghunna*, and *hamza*. We found the best Unicode-based font for properly displaying Urdu is Urdu Naskh Asiatype, available from the BBC Urdu Web site, at least among free fonts.

We compared this font (presumably optimized for Urdu) and Arial Unicode MS (presumably not optimized for Urdu) and found that the letter *heh* and its variations are mapped differently in these two Unicode-based fonts
(Table 1).

Table 1: How fonts display variations of the letter *heh*

	Urdu Naskh Asiatype display	Arial Unicode MS display
06C1 ہ	FBA6 ﮦ or FEE9 ﻩ FEEA ﻪ FBA8 ﮨ FBA9 ﮩ	FBA8 ﮨ FBA9 ﮩ
06BE ھ	FBAA ﮪ or FEEB ﻫ FEEC ﻬ	FBAA ﮪ or FEEB ﻫ
0647 ه	not found in corpus	FBA6 ﮦ or FEE9 ﻩ FEEA ﻪ FEEC ﻬ

For this reason, the metadata of the Urdu text in the corpus will contain the name of the Unicode-based font in which the text is stored. Any text processor that uses the data will have to normalize the usages of *heh* and its variations. In order to view the Urdu text properly in its surface form the font in which the data was harvested will have to be applied.

Differences in font mappings are not much of a problem when handling English and other Roman-based orthographies, especially when using the Unicode character set, so special attention has to be paid to the different ways fonts display surface forms of Urdu letters.

4 Urdu input method

In order to add an Urdu document to our corpus that we only have in graphic form or hard copy, we spent significant time setting up our computer for Urdu Unicode input in order to be able to type into the corpus.

Using the Arabic support on our computer, Microsoft Windows 2000 5.0 Service Pack 2, we were easily able to install right-to-left script support. Since Windows 2000 uses the Unicode character set internally, we did not have to do anything special to get Unicode support for our efforts.

Devising a plan for inputting Urdu on the keyboard was the biggest challenge. We ended up using Tavultesoft Keyman software to map our own keyboard—it was very easy to use. Existing keyboard mappings for Arabic script-based languages, we found, are generally not phonetically mapped, meaning we would like Urdu letter *feh* to be mapped to the letter *f* on the keyboard and so forth. We did find one phonetically mapped keyboard that we liked for Persian [3], CRL Phonetic Layout, so we used that mapping as a basis for developing our own. It is not important that our keyboard mapping be standardized—it only need work for the one person typing our text.

Conclusion

In this paper, we presented the methodology we used to build an Urdu corpus. The process of corpora construction for South Asian languages, specifically Urdu, involves extra work because these languages are not written in a Roman-based script. The use of the Unicode character set and software that supports it makes building needed corpora in these languages possible and relatively easy. Once corpora in these languages become readily available, natural language processing work in these languages can move forward.

References

[1] P. Baker and A. McEnery, “Needs of Language-Engineering Communities; Corpus Building and Translation Resources,” MILLE working paper 7, Lancaster University, 1999.

[2] N. Ide and G. Priest-Dorman, Eds., “Corpus Encoding Standard,” [Online document], 2000 March 20, [cited 2002 Feb 28], Available:
http://www.cs.vassar.edu/CES

[3] “Persian Keyboard Layouts,” Computing Research Laboratory, [Online document], [cited 2002 May 5], Available:
http://crl.nmsu.edu/~mleisher/keyboards/persian.html