A General-Purpose Monitor Corpus of Written Pashto

Mohammad Abid Khan[1]
and Fatima Tuz Zuhra[2]

Abstract

This paper provides an overview of the development of a general-purpose reference corpus for Pashto. It is an open-ended (monitor) corpus. The corpus currently contains 10,000 words. It has two cells, one containing essays and the other letters. This corpus represents Yousafzai group of dialects and the data for it has been provided by the Pashto Academy, Peshawar University. The corpus has been developed in Microsoft Visual Studio environment, with Extensible Markup Language (XML) at back-end and C# at front-end. The data in the XML form is tagged upto sentence level. The user-friendly front-end, in C#, allows a user to type a word in a text-box. The system searches the sentences containing this word and displays these sentences. This facility is useful for those language researchers who want to have example sentences of the use of a particular word. It also has a button for counting the total number of occurrences of a particular word, entered by the user, for searching in the corpus. This count is useful in situations where the users want to find out how frequently a particular word is used in the language. The whole interface is in Pashto language. The text in the corpus is handled in Unicode form. The paper provides an overview of how this corpus is useful for research purposes. This paper also provides the details of the methodology used for the development of this corpus. It has also been discussed that how different operations, such as searching the data, can be performed on the corpus using the interface provided by the corpus. Coding excerpts are given to facilitate the corpus development discussion.

1. Introduction

The definition of a corpus according to Sinclair (Sinclair, 2004: 19) is: "A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research".

The above definition reveals several facts about a corpus. First of all, a corpus is a collection of pieces of text. It means that a huge amount of text is available for a particular language. This text may be in the form of e.g. books, newspapers, journals, and periodicals. One of the properties of a corpus is that it is of a finite size. So, samples from huge population need to be taken. These samples are called pieces of language as is used in the above definition of corpus.

[1] Department of Computer Science, University of Peshawar
e-mail: m.abid6@gmail.com
[2] Department of Computer Science, University of Peshawar
e-mail: fateeshah@gmail.com

It has also been mentioned that the text, contained in a corpus, is in electronic form. It means that all the available text, such as hard covered books, needs to be transformed into machine readable form. For this purpose, any one of orthographic, prosodic, or phonetic transcription is implied.

Another fact from the definition of corpus is that the text in a corpus has been selected according to external criteria. It means that a domain must be selected, so that all the text, contained in the corpus, is according to that domain. For example, one may develop a spoken language corpus, a corpus of speeches recorded in academic environment, a learner's corpus, or a language acquisition corpus. Once the domain or the criteria has been selected, all the data for the corpus is gathered in accordance to it.

One of the points in the definition is that the text, having the above-mentioned properties, represents, as far as possible, a language or a language variety. It must be kept in mind that a corpus consists of samples, taken from the population of data. Representativeness means that these samples are taken in such a way that they represent all the possible features of a language. The sampling process must not cause important information to be removed from the text. The information lost during sampling process must be recoverable. Also, it means that all sorts of texts e.g. that of newspapers, novels, religious books and scientific text should be covered.

The most important point in the definition of the corpus is that, a corpus is used for linguistics research.

The study of language based on examples of 'real life' language use is called corpus linguistics (McEnery *et al.*, 2001: 1). This definition reveals that corpus linguistics is based on the language, used by humans, in real life rather than hypothetical examples. It has already been made clear that a corpus may consist of any type of data, including spoken language data. A corpus contains the language data as it is used by humans in daily life. Some of the sentences in a corpus may be ungrammatical, but are included in it because a native speaker of the language utters them.

Corpus linguistics deals with the study of language use. Different types of analyses may be applied on a corpus, based on the needs of a particular corpus linguist. For example, a corpus may be used as a tool to study the forms of a verb in Pashto (morphology); to study the hierarchical structure of the sentences and phrases of a language (syntax); or for discourse analysis. It is unpredictable during the corpus construction that for what purposes a corpus will be used. That is why, a corpus must be representative of a language, as far as possible.

Thus, corpus linguistics is the more scientific approach towards linguistics' study as it is based on facts, rather than assumptions.

The rapidly growing research work on Pashto language demands the construction of a corpus of the language. So far, no such corpus of machine readable text exists. Indeed a Pashto corpus has been developed for the BBN Byblos Pashto OCR System (Decerbo *et al.*, 2004), but this corpus contains Pashto data in the form of images. The data for the General-Purpose Monitor Corpus of Written Pashto, discussed here, contains data in Unicode form that is easily processable by any Unicode enabled application.

More and more data can be added to this corpus when it becomes available. Thus, the variation in the language with respect to time will be present. Currently, this corpus has 10,000 words. The corpus data is divided into two cells: one for essays and another for letters. All the text is stored in Arabic script.

The corpus facilitates the user to enter a query word or string in order to observe its different uses. This facility is needed by many researchers of the language. For instance, a language researcher may search the real uses of a word, he/she wants, rather than thinking of his/her own examples. The user can also count the total number of occurrences of a word or phrase in the corpus. This facility is useful in situations, where a researcher is interested in how frequently a word is used in the language.

2. The Development Methodology

The data for the corpus has been provided by the Pashto Academy, University of Peshawar. It is written Pashto data, containing essays and letters, written by different Pashto writers.

The tools used for the development of the corpus are XML at the back-end and C# at the front-end. The Arabic script has been enabled for the storage and retrieval of the Pashto data in Arabic script. A discussion on the tools used is given below.

2.1 Extensible Markup Language (XML)

Extensible Markup Language (XML) has been used at the back-end of the corpus. The data is stored in XML using tags. This tagging is of greatest interest for a corpus-developer. Text Encoding Initiative (TEI) standards can be enforced on a document easily using XML. UTF-8 encoding scheme is used in order to store the data in Arabic script. The corpus has two cells in the form of XML files: one contains the written essays and the other contains letters.

The main tag in the essay file is <essay>. It contains the sub-tags <title> i.e. the title of the essay, <author> i.e. the name of the writer of the essay, <date> i.e. the date of the publication of the essay, and <source> i.e. the source from which the essay has been taken. This document header is shown in Figure 1.

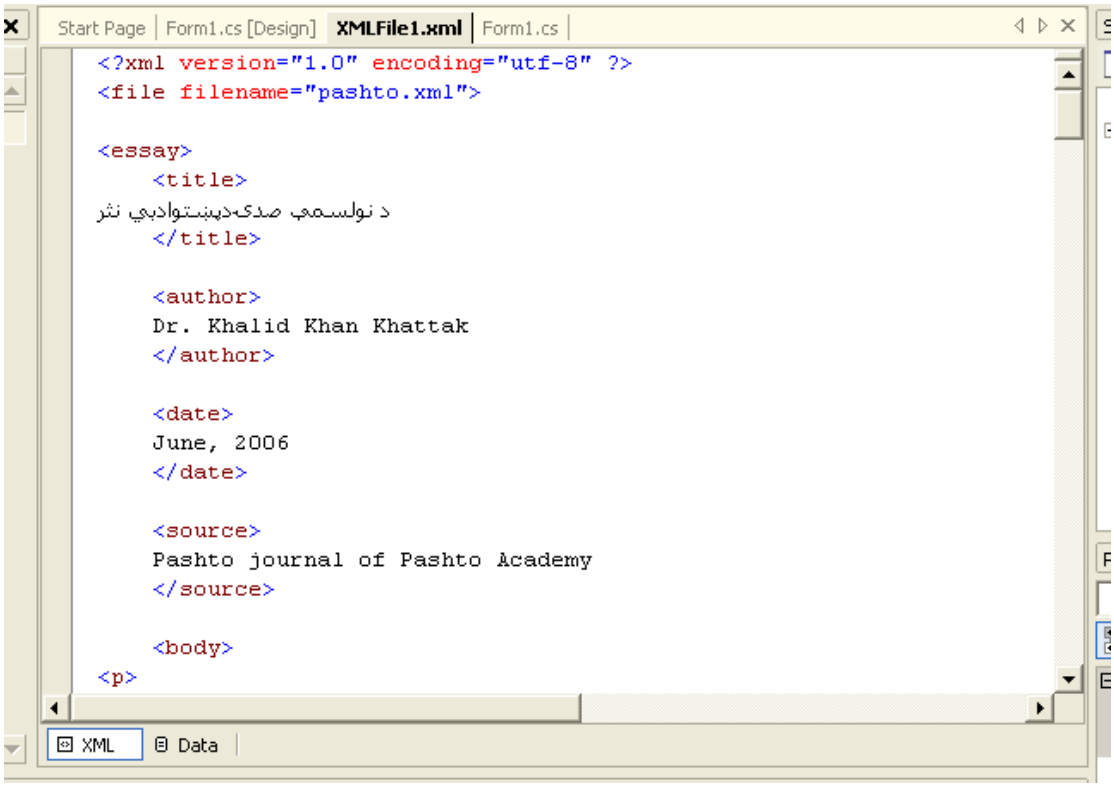

Figure 1: The header of the essays' section of the corpus

The <body> element of the essay contains the content of the essay. The body section is further tagged in paragraphs with <p> tag, and the sub-tagging of <p> in sentences with <s> tag. Figure 2 shows an example of this.

Figure 2: The body of the essays' section of the corpus

The letter section of the corpus contains the main tag <letter>. One of its sub-tags is <salutation> for the salutation portion. The letter body is divided into paragraphs, having tag <p> and the paragraphs are further divided into sentences, having tag <s>. Figure 3 shows the letter section of the corpus.

Figure 3: The letters' section of the corpus

The other sub-tags of the main tag <letter> are <signature> for signature portion, <date> for the date section, and <address> to show the address of the sender of the letter.

2.2 C#

The programming language C# is used at the front-end of the corpus. The user-interface will be discussed in the next section in detail. The files in XML form are loaded into C# as soon as the button for a particular search is clicked. The C# code searches the query word in both of the cells of the corpus and returns those sentences that contain the query word. These sentences are then displayed at the user interface to the corpus. The text displayed in this way is without tags. This is in accordance with what Leech (Leech, 2004: 25) and McEnery *et al*. (McEnery *et al*., 2001: 33) say. They say that if a corpus is tagged then there must be some way to hide these tags and to recover the original text. A sample from the C# code is shown in Figure 4.

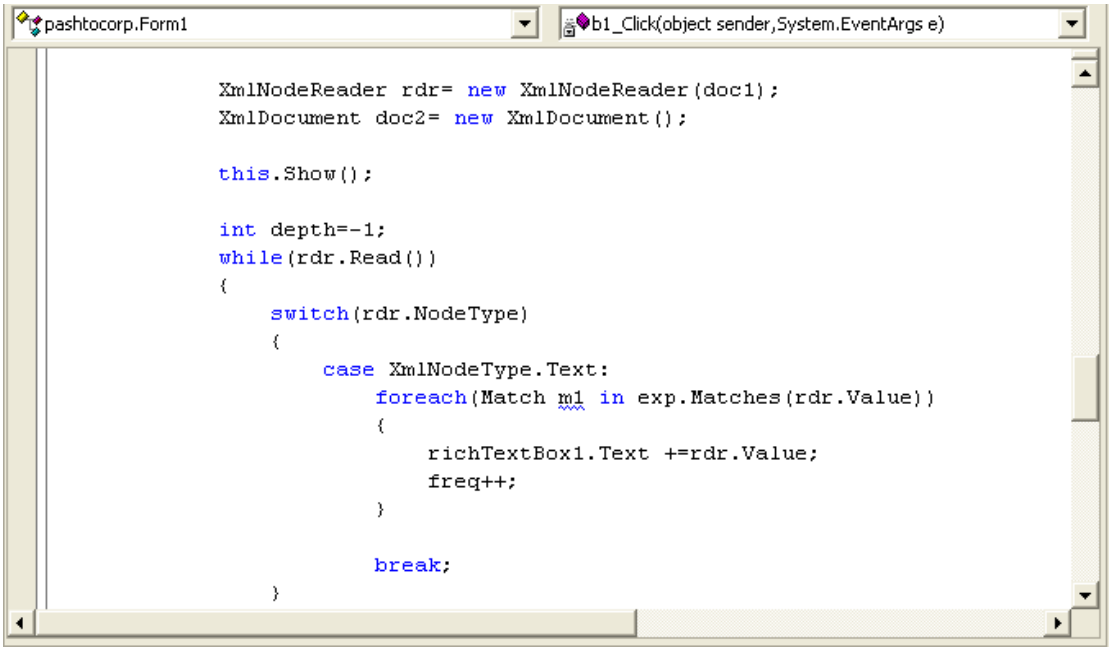

Figure 4: The coding excerpt

2.3 Algorithm

The sequence of steps i.e. the algorithm of the corpus is given below.

Step1. Present the user with the user interface.
Step2. If the user clicks لټون *button then do step 3 and onwards.*
Step3. Load the XML files in C#.
Step5. Take the text from the input textbox, add a hard space before and after it and save it in a variable of type string. This becomes the query text.
Step6. Search the query text in the loaded XML files.
Step7. If match is found then do step 8
else do step10 and onwards.
Step8. If the output textbox already contains the sentence having the query word, then do step 9
else print the sentence having the query word, in the output textbox and do step 9.
Step9. Increment the frequency counter and do step 7.
Step10.If شمېر *button is clicked then display the frequency count*
else check whether the user clicks لټون نوی *button*
If yes then clear the text from all textboxes and do step 2
else Exit.

2.4 Flowchart

The flowchart in Figure 5 provides details of the working of the program.

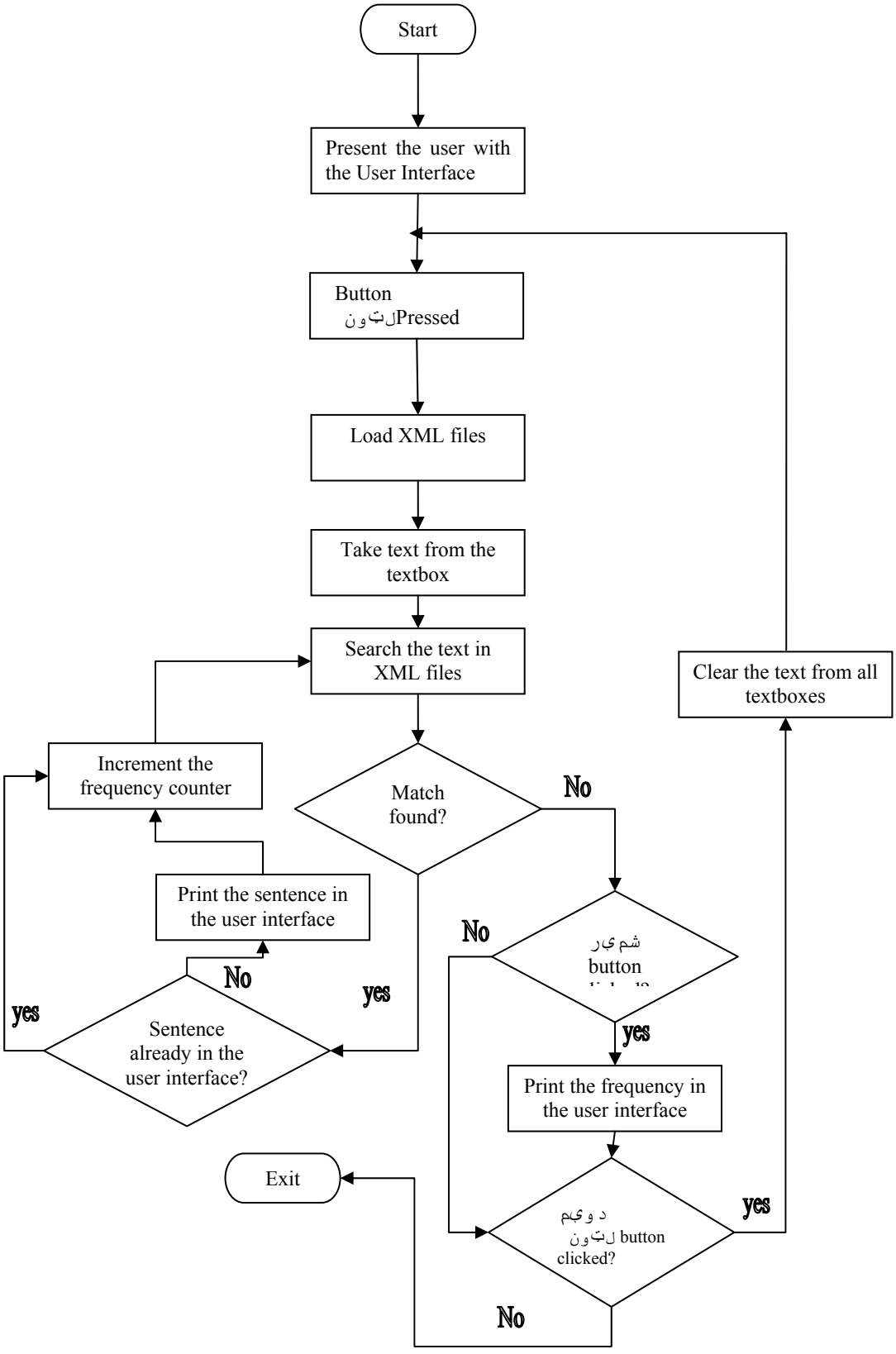

Figure 5: Flowchart of the system

3. Interface to the Pashto Corpus

The corpus has a user friendly interface. The whole interface is in Pashto, with its right-to-left script. It allows the user to enter a query word in a textbox, labeled with a message to the user to enter a word. It is shown in Figure 6 below.

Figure 6: The user interface of the corpus

When the user enters a word and clicks the button لټون, the sentence(s), containing the query word are displayed in the output rich-textbox. It is shown in Figure 7.

Figure 7: The data searched in the corpus

There can be a problem in a corpus similar to this corpus. The problem is that if some word is repeated two or more times in a sentence, then the corpus interface may display the sentence as many times as the word is repeated. For example, the word "دا" comes two times in "ما سره خو دا ويره ده چرته دا هر څۀ مو په اوبو لاهو نۀ شي ـ". An ordinary corpus interface may display this sentence two times. The General-Purpose Monitor Corpus of Written Pashto does not encounter this problem.

If the query word is not found in the corpus, then the corpus interface simply displays an informative message, in red color, to the user that the query word is not found in the corpus rather than being crashed or hanged. It is shown in Figure 8.

Figure 8: The corpus interface response on a word not found the corpus

There is a button, named شمېر that allows a user to see the number of occurrences of the query word in the corpus. An example of the use of this button is shown in Figure 9.

Figure 9: The frequency count displayed by the corpus

Another button, د ویم لټون, is used to refresh i.e. to clear the textboxes at the user interface if the user wants to do another corpus search.

4. Conclusion

In this paper, the development of the General-Purpose Monitor Corpus of Written Pashto has been discussed. The corpus is tested and is successful. This paper also discusses how a corpus, especially a Pashto corpus can be useful in the study of the language.

References

Decerbo, M. *et al*. (2004) 'The BBN Byblos Pashto OCR System'. *ACM Press*.

Leech, G. (2004) Adding Linguistic Annotation, in M. Wynne (ed.), Developing Linguistic Corpora: a Guide to Good Practice, pp. 21–36. University of Oxford: AHDS Literature, Languages and Linguistics.

McEnery, T. *et al*. (2001) Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University Press.

Sinclair, J. (2004) Corpus and Text-Basic Principles, in M. Wynne (ed.), Developing Linguistic Corpora: a Guide to Good Practice, pp. 1–20. University of Oxford: AHDS Literature, Languages and Linguistics.

TEI Guidelines available on-line from http://www.tei-c.org/release/doc/tei-p5-doc/html/ (accessed 10th December 2006)