1. Installation and execution

1.1. For Windows

• Unzip source.tar.gz and open this directory in command prompt.
• To compile the source, type compile.bat. (The compilation requires the Haskell GHC compiler, a C compiler and Java SE Development Kit (JDK) 5.0)
• To run application, type: run.bat

1.2. For Unix/Linux/Solaris

• Unzip source.tar.gz and open this directory in shell.
• To compile the source, type ‘make’. (The compilation requires the Haskell GHC compiler, a C compiler and Java SE Development Kit (JDK) 5.0)
• To run application, first type: chmod a+x run.sh and then type: ./run.sh

Haskell GHC compiler could be taken from (http://www.haskell.org/ghc/)
A gcc C compiler for windows could be taken from (http://www.delorie.com/djgpp/)

2. The available modes and tools in the Application

The Java GUI application interfaces the FM runtime system to provide morphological analysis both in Urdu and Roman. We Interface four kinds of analysis which are provided by FM as a part of its runtime system. They could be invoked from the Menu “Analysis”. In each analysis mode, a screen keyboard is provided to let a user type Urdu.

We now discuss the analysis modes below one by one:

2.1. Tagger Mode:

This mode analyses the given words into their dictionary form and displays their grammatical description.

Let us tag some words, (کتاب, book) and (پڑھانا, to teach)
Dictionary loaded: DF = 4131 and WF = 496976.

[<ةَبَطاَبٰ>, ktab>
1. {ةَبَطاَبٰ, ktab} (802) Noun - Singular Nominative - Feminine
2. {ةَبَطاَبٰ, ktab} (802) Noun - Singular Oblique - Feminine
3. {ةَبَطاَبٰ, ktab} (802) Noun - Singular Vocative - Feminine
]

[<پڑھانا>, pR|hana>
1. {پڑھانا, pR|hna} (320) Verb - Caus1 Inf -
]

The DF count is the number of dictionary words in the lexicon, and WF count is the number of word forms. Since there are a big number of word forms (WF) for this lexicon, Tagger Mode might take some time to build the data structure for the first run.

2.2. **C-Trie program:**

It provides similar but faster analysis as compare to Tagger Mode. It is implemented in C language as a part of FM. This mode uses urdu.fullform lexicon file.

The analysis for the same words (کتاب, كتاب) and (پڑھانا, پڑھانا) is following:

```
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Trie 0.1, April 2004</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Author: Markus Forsberg</td>
</tr>
<tr>
<td>Send bug reports to: <a href="mailto:markus@cs.chalmers.se">markus@cs.chalmers.se</a></td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Interfaced for Urdu</td>
</tr>
<tr>
<td>by Muhammad Humayoun 2006</td>
</tr>
<tr>
<td>Chalmers University of Technology</td>
</tr>
<tr>
<td>-----------------------------------------------</td>
</tr>
</tbody>
</table>
[Reading lexicon read from ‘urdu.fullform’...]  
[ Number of entries :  662k entries ]
[ Build CPU time :  8.35 seconds ]
```
ktab:
1. كتاب, ktab (799) Noun - Singular Vocative - Feminine (0)
2. كتاب, ktab (799) Noun - Singular Oblique - Feminine (0)
3. كتاب, ktab (799) Noun - Singular Nominative - Feminine (0)

pR\hana:
1. پڑھنا. pR\hana (314) Verb - Caus1 Inf - (0)

2.3. Synthesiser Mode:

The synthesiser mode takes a word form and generates the complete inflection table of the group from which the word belongs.

The analysis for word (كتاب, کتاب) is shown below:

************************************************************************************************************
*    Urdu Morphology                                    *
************************************************************************************************************
* Functional Morphology v1.10                          *
* (c) Markus Forsberg & Aarne Ranta 2004               *
* under GNU General Public License.                   *
************************************************************************************************************
* Implementation for Urdu as a Master Thesis           *
* (Muhammad Humayoun 2006)                            *
* Chalmers University of Technology                    *
************************************************************************************************************

Dictionary loaded: DF = 4131 and WF = 496976.
[Synthesiser mode]

ktab
Noun
Feminine
Singular Nominative: ktab, كتاب
Singular Oblique: ktab, كتاب
Singular Ergative: ktab-nE, كتاب-نی
c
Singular Accusative: ktab-kw, كتاب-کو
c
Singular Dative: ktab-kw, كتاب-کو
c
Singular Instrumental: ktab-sE, كتاب-سی
c
Plural Nominative: ktabyN, كتابین
c
Plural Oblique: ktabyN, كتابین
c
Plural Ergative: ktabyN-nE, كتابین-نی
c

Like Tagger mode, Synthesiser mode might also take some time to build the data structure for the first run.
2.4. Inflection Mode:

The inflection mode takes a word form and an interface function defined in CommandsUrdu.hs (such as n1, n2 …for nouns) and either the word exists in the lexicon or not, it displays a complete inflection table for that word form. The inflection mode is the same as synthesis mode with no lexicon loaded. This mode could be used to learn the inflections of a certain paradigm.

A user requires to select command (n1, n2, v1, v2 etc) from the combo box and type the word in the text box to get the inflections. To demonstrate it, suppose we select command n1 and type a word (mʊnɖɑ, منّدا, boy) which is a Punjabi word and definitely could not be the part of Urdu lexicon.

```
Dictionary loaded: DF = 4131 and WF = 496976.
[Inflection mode]
```

m(o)nDa
Noun
Masculine
Singular Nominative: m(o)nDa, : مّنّدا
Singular Oblique: m(o)nDE, : مّنّدے
Singular Ergative: m(o)nDE-nE, : مّنّدے-نے
c Singular Accusative: m(o)nDE-kw, : مّنّدے-کو

2.5. Urdu Transliterator

It is a utility for the conversion of Urdu to Roman and Roman to Urdu script provided in Utilities menu.

2.6. Unicode Urdu Extraction

It is a utility for the extraction of Urdu text from Web pages or text pages.
This utility takes a web/local address, removes all HTML tags and returns the actual text written on that page. It also removes the text other than Urdu Unicode text, only permitting Urdu Unicode characters. Therefore if a web page contains text written in ASCII Roman (e.g. English) and in Unicode Urdu, only Urdu text is returned. Then the extracted Urdu could be saved in appropriate format by using this utility.

3. Some issues

3.1. The Runtime memory issues

For Tagger and synthesiser mode a user could expect the following error message that states that the memory assigned to the FM Runtime is not sufficient for the morphological analysis.

.....
.....
Current maximum heap size is 268435456 bytes (256 MB);
use +RTS –M<size> to increase it.
Process Ended....

In such cases a command “memory” is provided under Settings menu in the Application. A user must add the appropriate memory for the Runtime system if needed. A typical value could be 350 (350 MB).

3.2. The Path issues

Please copy the application on such path where space does not occur in the directory names. For example copying it on a path “C:\Documents and Settings\Administrator\Desktop” is not good as “Documents and Settings” directory contains space characters. While copying it on a path like “D:\docs\WorkSpace\abc\xyx\project” is perfectly fine.