Turku Izhevsk Corpus v1.1

The corpus was created by the Research Unit for Volgaic Languages (University of Turku) in collaboration with the Language Department of the Udmurt Institute of History, Language and Literature (Izhevsk).

The texts were prepared for the corpus by Jorma Luutonen (Turku) and Leonid Ivshin (Izhevsk).

The contents of the corpus

The corpus contains ca. 11 000 texts from a newspaper and five journals:

Udmurt dunne 10 366 texts (years 1997,1998,1999, 2000, 2001)
Dzhetshbur 152 texts
Vordskem kyl 139 texts
Invozho 130 texts
Kenesh 116 texts
Kizili 116 texts

Transliteration and file names

The corpus texts were originally transliterated from Cyrillic to Latin alphabet. The alphabet was later mechanically changed back to Cyrillic, which caused some regular errors that have not been corrected. For instance, Roman numbers appear as И, ИИ ИИИ, ИВ, В, etc. A fixed line number has been inserted to the beginning of each line.

The names of the text files are usually numbers, e.g. 1.txt, 2.txt, 3.txt, etc. The text files were originally located in folders. The "Nice name" of a corpus text in the Finno-Ugric Corpora portal version of the corpus reflects the old folder division (see next section), for instance the nice names Invozho/C/16, and Udmurt_dunne/1998/025/16.

The old subfolders

The original grouping of the texts by the editors of the newspaper has generally been preserved in the case of Udmurt dunne. Each annual volume has been subdivided into folders which contain the texts. The folders and subfolders are as follows:

```
Udmurt_dunne
   1997
       119....162
   1998
       002...163
   1999
```

```
        001...147
    2000
        001...160
    2001
        001_3...048-49
```

In the case of the journals, the texts are located in subfolders labelled A, B, C, etc. These subfolders generally correspond to those floppy disks in which the material was obtained from the editorial offices of the journals. The folders and subfolders include the following:

```
Dzhetshbur
   A...C
Invozho
   A...I
Kenesh
   A...H
Kizili
   A...H
Vordskem_kyl
   A...J
```

It is not known how the subfolders correspond to the printed numbers of the newspaper or the journals. However, it is reasonable to expect that the subfolders of Udmurt dunne, at least to some extent, correspond to actual issues of the newspaper. For the creation dates of the text files from the journals, see next section.

Creation dates of the journal text files

It is not known which number of a journal a corpus text belongs to. The following information about the creation dates of the text files might be useful if one wants to locate the texts.

DZHETSHBUR

The names of the text files are original, but the meaning of the numbers in the file names has not been clarified.

```
A  :  files created 3-4/2001
B  :  files created 2-3/2001, 10/2001 and 3-4/2002
C  :  files created 4-8/2001
```

INVOZHO

```
A :   some of the files were created 1/2001
B :   files created 2/2001
C :   files created  4/2000, 7-8/2000, 10-11/2000, 2/2001, 4/2001
D :   files created 3-4/2001, 6/2001
E :   files created 7-9/2001
F :   files created  10-11/ 2001
G :   some of the files were created 12/2001 and 1/2002
H :   files created 1-2/2002
I :   files created  2-3/2002
```

KENESH

(In many cases, the original creation date is not known.)

A : the text files belong to the issue 4/2001; files created 3-4/2001
B : the text files belong to the issues 5/2001, 6/2001, 11/2001 and 12/2001; some of the files were created 3-5/2001
C : the text files belong to the issues 7/2001, 8/2001 and 9/2001; some of the files were created 5-7/2001
D : the text files belong to the issues 7/2001, 8/2001 and 9/2001; the files were created 3-7/2001
E : creation date of one file known: 3/2002
F : creation dates of some files 3-4/2002
G : (creation dates not known)
H : (creation dates not known)

KIZILI

A : creation dates of the files 7/1997, 9/1997 and 10/1999
B : files created 5-6/1999
C : files created 5/2001
D : files created 3-5/2001
E : files created 5-7/2001
F : files created 7/2001
G : most of the files were created 4/2001, one file 5/2001 and another 7/2001
H : most of the files were created 4/2002

VORDSKEM KYL

A : files created 12/2000, 1-2/2001
B : files created 2-3/2001
C : files created mostly 2-4/2001
D : files created 3-6/2001
E : files created mostly 6-7/2001
F : files created 10/2001
G : files created 9-10/2001
H : files created 11/2001
I : files created 1-2/2002
J : files created 2-3/2002

References to the corpus

As it is not known in which issue of a newspaper or a journal a certain text has been published, referring to the corpus itself is safest. For uniformity, the references should have the following form:

Turku Izhevsk Corpus/Udmurt dunne/1999/21/5:33
(Line 33 in the text 5 of the subfolder 21 of the year 1999 of Udmurt dunne.)

Turku Izhevsk Corpus/Kenesh/C/11:17
(Line 17 of the text 11 in the Kenesh subfolder C.)

If necessary, the Turku-Izhevsk Corpus can be abbreviated TIC, in which case the references appear as:

TIC/Kenesh/C/11:17.