Quick Start Manual

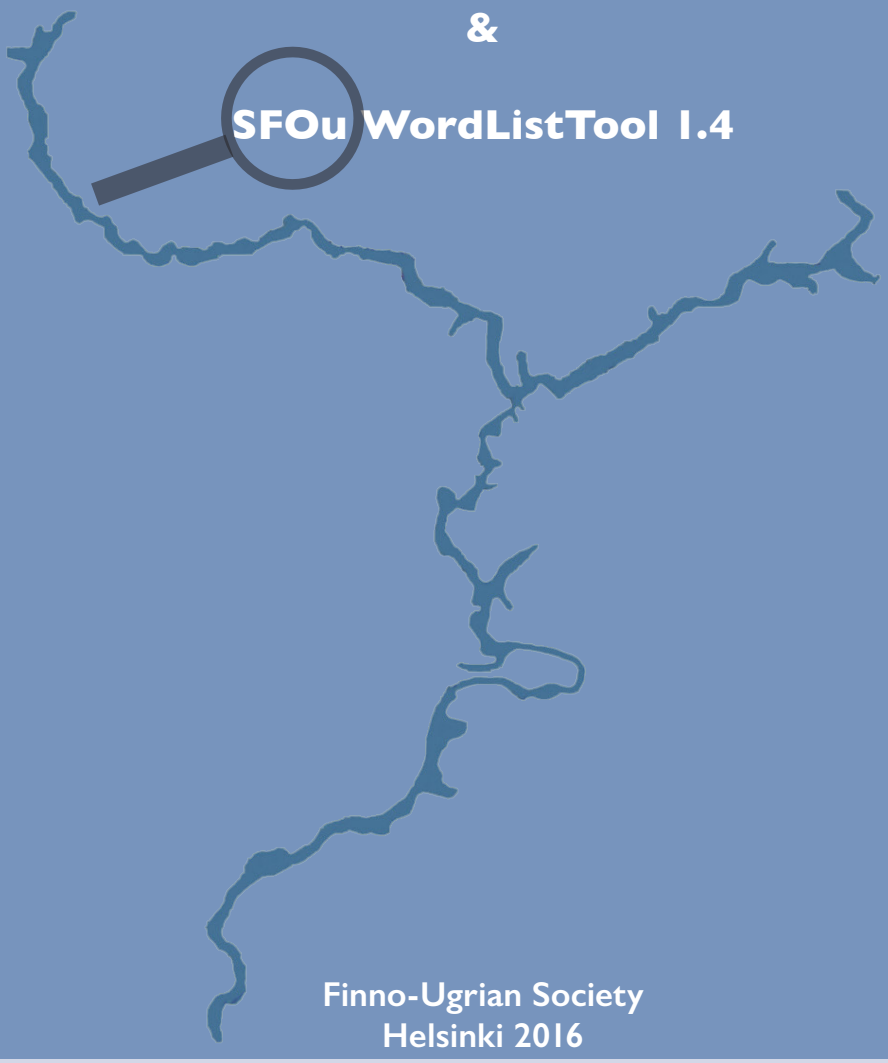
for

Jorma Luutonen et al.

Electronic Word Lists: Komi, Chuvash and Tatar
Lexica Societatis Fenno-Ugricae XXXI:2

&

SFOu WordListTool 1.4



Finno-Ugrian Society
Helsinki 2016

The publication includes

Enye Lav and Jorma Luutonen: *An Electronic Word List of Komi*
Pavel Zheltov, Eduard Fomin and Jorma Luutonen: *An Electronic Word List of Chuvash*
Mansur Saykhunov and Jorma Luutonen: *An Electronic Word List of Tatar*

&

A special user interface program *SFOu WordListTool 1.4*

Electronic Word Lists: Komi, Chuvash and Tatar / Электронные списки слов: коми-зырянский, чувашский и татарский языки / Sähköisiä sanalistoja: komi, tšuvassi ja tataari
Jorma Luutonen et al. (Jorma Luutonen, Eduard Fomin, Enye Lav, Mansur Saykhunov, Pavel Zheltov)
Lexica Societatis Fenno-Ugricae XXXI:2
Helsinki 2016
Layout / вёрстка / taitto: Eeva Herrala
Cover / обложка / kansi: Eeva Herrala, Anna Kurvinen

ISBN 978-952-5667-79-0
ISSN 0356-5769

SFOu WordListTool 1.4

Contents

1. Introduction

In the past, large word lists with word class labels have been published by the Finno-Ugrian Society in the form of reverse dictionaries. The languages represented in these publications are Mari (2002), Mordvin (2004), Chuvash (2009) and Komi (2012); for detailed information, see references. The main purpose of reverse-alphabetised word lists is to serve as a source for the study of derivation and word structure.

In 2007, the vocabularies of the Mari and Mordvin reverse dictionaries, together with a word list of Udmurt, were published as *Electronic word lists: Mari, Mordvin and Udmurt* (Lexica Societatis Fenno-Ugricae XXXI:1). The word list of Udmurt was mainly based on the material of an Udmurt reverse dictionary printed in Izhevsk in 1992.

The present collection of electronic word lists contains lexical items from the aforementioned Chuvash and Komi reverse dictionaries as well as a word list of Tatar created especially for this publication. Known errors have been corrected in the Chuvash and Komi word lists, and they have been supplemented with new words. In the case of Chuvash, the number of new entries, 139 out of a total of 31,403, is relatively small. The Komi word list, however, was substantially enlarged: there are 22,544 new items, which make up a third of the total number of 70,199 entries in the list. The Tatar word list, published for the first time, is essentially an edited combination of the vocabularies of three recently published Tatar dictionaries. Detailed descriptions of all three word lists can be found in the file *Descriptions_en_2016.pdf*, which is located in the *documentation* subfolder of the SFOu WordListTool program folder.

The user interface program *SFOu WordListTool*, produced for the first electronic word list publication in 2007 by Turku University of Applied Sciences, has been updated for the present publication. (The abbreviation SFOu refers to the French name of the Finno-Ugrian Society.) The new version 1.4 contains new features that make the program more flexible and user-friendly. For more information, see *User_guide_en.pdf* in the *documentation* subfolder of the SFOu WordListTool program folder.

2. Installation

The word list files can be used with word processors that can cope with Unicode characters, or with the help of such programs as Microsoft Excel. It is, however, suggested that the user utilises the *SFOu WordListTool* program that has been developed specifically for these kinds of word lists.

There are installers for *SFOu WordListTool* and the word lists for PC and Mac computers. For PCs, Windows 2000/XP/Vista/7/8, and for Macs, Intel based Mac OS 10.6 or newer is required. PowerPC based Macs should use the old 1.3 version of the program. You can install the program and the word lists by following the instructions below.

Installation on Windows OS

1) If the installation doesn't start automatically, double-click on the ***SFOuWordListTool.exe*** file.
2) From the window which appears, select Setup Language and click 'OK' to continue the installation.
3) Setup Wizard welcomes you to the SFOu WordListTool installation, click 'Next' to continue.
4) Read through the Licence Agreement and click 'I accept the agreement' if you accept the terms of the agreement. Click 'Next' to continue.
5) Select the Destination Location, where the program should be installed. Click 'Browse' to choose a custom folder. Click 'Next' to continue the installation into the selected destination.
6) Select Start Menu Folder. Click 'Browse' to choose custom folder. Click 'Next' to continue the installation into the selected destination. If you do not want to create a Start Menu folder, tick the 'Don't create a Start Menu folder' box and then click 'Next' to continue.

7) Select Additional Tasks. Tick the additional boxes if you want to create a Desktop icon or a Quick Launch icon. Click 'Next' to continue.
8) In the following window you will see a summary of installation settings. Click 'Install' to continue with the installation or click 'Back' if you want to review or change any settings. Click 'Cancel' to terminate the installation.
9) Setup has now finished installing the SFOu WordListTool on your computer. Tick the 'Launch SFOu WordListTool' box if you want to launch the program immediately after installation. Click 'Finish' to exit Setup.

Installation on Mac OS

1) Drag the SFOu WordListTool from the opening Finder window to the Applications directory or another directory where you wish to access the program. (If you do not have the privileges required to install the program in the intended directory, the operating system will ask for the administrator's password. If the required privileges are not available, the program can only be installed in your home directory.)
2) If you want to create a shortcut for the program in the Dock bar, you must
    a) go to the installation directory, and
    b) drag the SFOu WordListTool.app icon to the Dock bar.

Once the program and the word lists have been installed, launch the program and read the instructions in the Help menu.

3. The word list files and their general structure

The word list package includes two files for each language, one with normal alphabetisation (beginning from the first letter of the word) and the other with reverse alphabetical order (beginning from the end of the word). The names of the files are as follows:

komi_alph.txt,
komi_rev.txt,
chuvash_alph.txt,
chuvash_rev.txt,
tatar_alph.txt,
tatar_rev.txt.

These files can be found in the *wordlists* subfolder of the program folder.

The character encoding of the files is Unicode (UTF-8). In the files, the material is arranged in four columns:

1) the word
2) language
3) word class
4) sources

The meanings of the words are not given in the word list. Technically, the files are plain text Comma Separated Value (CSV) files. This simply means that a comma character (,) separates the fields for different types of information (word, language, word class, sources) in each line of the file.

4. Summary description of the columns

The following sections briefly summarise the main characteristics of the lists and the abbreviations used to encode the information in them. For full descriptions, see the file *Descriptions_en_2016.pdf* in the package.

4.1. The first column: The word

All words are given in Cyrillic letters. The stress is marked with a dot (·) in a couple of words in the Chuvash word list. White spaces between the separate elements of a compound word have been substituted with low line characters (_), which may be invisible if you use the *SFOu WordListTool*. The following alphabetical order is used in the word lists:

А Ӑ Ә Б В Г Д Е Ё Ӗ Ж Ӂ З И І Й К Л М Н Ң О Ӧ Ө П Р С Ҫ Т У Ӳ Ү Ф Х Һ Ц Ч Ш Щ Ъ Ы Ь Э Ю Я

Stress marks "·", hyphens "-", low line characters "_" and brackets "(", ")" are ignored in the alphabetisation. Note that the alphabet the program uses for alphabetisation includes Latin characters as well as special Cyrillic letters used in writing the languages of the 2007 word list publication, Mari, Mordvin and Udmurt.

4.2. The second column: Language

These are the meanings of the abbreviations:

kom = Komi
chu = Chuvash
tat = Tatar

4.3. The third column: Word class

These are the abbreviations for the word classes, in alphabetical order:

ad = adjective
av = adverb
co = conjunction
de = descriptive word
ge = gerund (not used in the Tatar word list)
in = interjection (incl. words used in calling animals)
no = noun
nm = numeral
pa = particle
po = postposition
pc = participle (not used in the Tatar word list)
pr = pronoun
st = state (used only in the Komi word list)
vb = verb

Alternative word classes are separated by a slash (/), e.g. **ad/av** meaning adjective or adverb.

4.4. The fourth column: Sources

Full bibliographical data are given in the file *Descriptions_en_2016.pdf* in the *documentation* folder. The names of some sources have been abbreviated for clarity.

4.4.1. The Komi word list

These are the abbreviations for the sources of Komi words:

a = Коми-роч кывчукӧр ed. by Beznosikova (2000);
c = Русско-зырянскій словарь by Tsember (1910);
e = The lexical database of the Komi Zyryan spellchecker (2013);
E = Комиын овъяс by Plesovskiy (1997);
f = Syrjänisches Wörterbuch by Fokos-Fuchs (1959);
h = Краткий коми-русский словарь by Shakhov (1924);
i = Комі орфографіја кывӧктӧд by Razmanov (1930);
K = Jypidьіческӧj da кантселъарскӧj ъерминъяс (1924);
L = Коми пемӧс нимкуд by Tsypanov (2008);
m = Vocabularium Harmonicum by Müller (1759);
N = Collections of neologisms (Информационный бюллетень1994-2003; Выль коми кыввор 1998, 1999; Выль кыввор 2007);
o = Коми орфография кывкуд ed. by Karmanova (2008);
p = Коми-русский словарь ed. by Podorova (1948);
r = Русско-коми словарь ed. by Beznosikova (2003);
s = Зырянско-русскій и русско-зырянскій словарь by Savvaitov (1850);
t = Комиа-роча словарь ed. by Lytkin (1961);
u = Syrjänischer Wortschatz byWichmann (1942);
V = Краткий коми-русский, русско-коми ботанический словарь by Rakin (1989).

4.4.2. The Chuvash word list

These are the abbreviations for the sources of Chuvash words:

a = Чăваш терминологийĕ by Andreyev and Danilova (2001);
b = Инновации в лексике церковно-богослужебных текстов by Studentsov (2007);
c = Чувашско-русский словарь ed. by Skvortsov (1982);
d = Чăваш чĕлхин çĕнĕлĕх словарĕ by Degtyarev (2003);
e = Школьный русско-чувашский словарь: математика, физика, астрономия by Yelkin et al. (1996);
f = Диалектологический словарь чувашского языка by Sergeyev (1968);
g = Авалхи грек тата латин терминоэлеменчĕсен словарĕ by Fomin (2007);
h = personal observations of E. V. Fomin (for Russian translations of these words, see the file *Chu_source_h.pdf*);
i = Медицинский терминологический русско-чувашский словарь by Ivanov and Minnebayev (1998);
j = Русско-чувашский словарь юридических терминов by Skvortsov and Semenov (2006);
k = Ял хуçалăх терминĕсен вырăсла-чăвашла словарĕ by Kostin and Ignatyev (1976);
l = Авалхи халал: Моисейĕн пилĕк кĕнеки (2001);
m = Русско-чувашский словарь медицинских терминов by Grigoryev (1996);
n = Язык современного чувашского социума by Degtyarev (1999);
o = Чăваш чĕлхин орфографи словарĕ comp. by Alekseyev (2002);
p = Краткий русско-чувашский психиатрический словарь by Golenkov and Dolgova (2000);

q = Чăваш халăх сăмахлăхĕ (1973 – 1987);

r = Русско-чувашский общественно-политический словарь comp. by Skvortsov (1972);

s = Именные композиты в чувашском языке by Semenova (2002);

t = Русско-чувашский словарь технических терминов by Petrov (1971);

u = Чăваш чĕлхин ӳнер терминĕсем by Degtyarev (2007);

v = Чувашско-русский толковый словарь названий лиц by Sergeyev (2004);

w = Лесной словарь by Entip (1998);

x = Словарь чувашского языка by Ashmarin (1994 – 2000 (1928 – 1950));

y = Словари чувашско-марийских и марийско-чувашских заимствований by Fedotov (1990);

z = Русско-чувашский словарь социальной лексики by Skvortsov (2004).

4.4.3. The Tatar word list

These are the abbreviations for the sources of Tatar words:

c = Corpus of written Tatar (2013);

g = Татарская грамматика II ed. by Zakiyev et al. (1997 (1993));

h = Татар теле морфологиясе by Khisamova (2006);

k = Татар теленең орфографик сүзлеге ed. by Galiullin and Raskulova (2010);

m = Personal observations of M. Saykhunov;

o = Татар теленең орфография сүзлеге by Ganiev and Sabitova (2002);

r = Татарча-русча сүзлек ed. by Asylgarayev et al. (2007);

t = Татарская грамматика I ed. by Zakiyev et al. (1995 (1993)).

5. Differences between the word lists

The package *Electronic Word Lists: Komi, Chuvash and Tatar* consists of three large vocabularies, each of which has its own compilation and editing history. Although there was an explicit effort to maintain uniformity during the preparation of the lists, the final products differ from each other in some respects. Full descriptions of the characteristics of each list can be found in the file *Descriptions_en_2016.pdf* in the *documentation* folder. Some specific characteristics of each word list are briefly explained in the following sections.

5.1. The Komi word list

The Komi word list is the largest of the three in the package and it also covers a longer period – from the 18th century to the present day – than the two other lists. The inevitable heterogeneity of material originating in different periods of the development of the literary language has been reduced through phonetic and orthographic unification in accordance with the current (2013) rules of orthography. This means that the spelling of a word in the present electronic word list may differ from that in the lexicographical source. The user should also note that some old lexemes are obsolete, and many of the neologisms found in the list are theoretical creations that will be unfamiliar to the majority of Komi speakers.

Unlike the other two word lists, the Komi list contains a large number of proper nouns (place names, person names) systematically collected for the list.

A proportionally greater number of new Russian loanwords have been included in the Komi word list than in the Chuvash and Tatar lists. One reason for this is that the two sources used to increase the number of entries in the last phase of the work, the Komi spellchecker's database and the Russian-Komi dictionary (2003), contain many Russian loanwords.

As for compound words, the Komi word list only includes those written as continuous sequences of graphemes or where the components of a compound are connected to each other by a hyphen (-). If the parts of a compound are written separately, the word has not been included in the list.

The word class "state" (категория состояния), which comes from the Russian grammatical tradition, has been used in the Komi word classification. It refers to words expressing the state of an entity, such as the word *яндзим* 'shame' in the expression *сылы яндзим* 'he/she is ashamed'; literally: '(it is) shame for him/her', cf. Russian *стыдно* 'it is a shame' in *ему стыдно* 'he is ashamed'.

5.2. The Chuvash word list

In the Chuvash word list, the position of the stress is indicated by a dot (·) after the stressed vowel in those cases where it is only the stress that distinguishes two homographs, e.g. *каска·* 'log', *ка·ска* 'helmet'.

The word list contains over 300 compound words that consist of two parts separated by a low line character (_), which corresponds to the white space character in normal writing. Most of these compounds are written together as one word in the basic vocabulary source (**c**) of the list, *Чувашско-русский словарь* (1982), which represents an older orthographical convention than that followed by the compilers of the present word list.

The Russian loanwords found in the basic vocabulary source (**c**) were included in the list, but only a restricted number of new Russian loans have been selected from other sources.

Fewer words are classified as postpositions in the Chuvash word list than in the Komi and Tatar lists. The reason is that, in accordance with the Chuvash grammatical tradition, case forms of auxiliary nominal stems denoting spatial relations, such as 'front', 'back', 'up', 'down', e.g. *ум* 'front part' > *умĕнче* 'in front of (him/her/it)', *умĕнчен* 'from the front of (him/her/it)', were regarded as inflectional forms of nouns and thus omitted from the list.

5.3. The Tatar word list

The Tatar word list is essentially based on three general dictionaries, which means that the number of special terms used in different fields of knowledge is probably proportionally smaller than in the Komi and Chuvash word lists. There was also a conscious effort to restrict the number of new Russian loanwords when compiling the word list.

Unlike the two other word lists in the package, no orthographical changes have been made to the lexical items. An important exception, however, is that there are no capital letters in the word list, which means that all proper nouns begin with a small letter, e.g. *татарстан* 'Tatarstan'.

An overwhelming majority of the over 500 cases where the parts of a compound are written separately (connected by a low line character in the list) consists of verbs such as *харап булу* 'to get into big trouble, to die'.

The categories gerund (**ge**) and participle (**pc**) are not used in the classification of Tatar words. Lexicalised gerunds are labelled as adverbs (**av**) and lexicalised participles as adjectives (**ad**). There are many fewer alternative classifications of words (e.g. **no/vb**) than in the Komi and Chuvash word lists.

6. Some practical advice for the user

The most difficult task in the compilation process was determining word classes for the words. In the case of the smaller categories (other than verb, noun, adjective), there undoubtedly remain inconsistencies. The information in the word class column is not intended to give a definitive answer to the classification of every word; rather it is provided in order to help the user, to give him or her at least one possible classification. In unclear cases, the user is advised to look at the corresponding entries in the dictionary sources.

You can compare languages by doing searches simultaneously from two or three word lists that represent different languages. When using the *SFOu WordListTool*, you can choose the way in which the search results will be alphabetised from the settings tab.

The *SFOu WordListTool* program allows the user to define search objects using formulas called *regular expressions*. Half an hour spent in learning the basics of regular expressions will probably save you a great deal of time in the future. Easy introductions to regular expressions can be found in the Internet, e.g. in Wikipedia.

In the *SFOu WordListTool*, you can re-alphabetise an already existing word list file by giving .+ (i.e. dot followed by plus) as the string to be searched and choosing the alphabetisation mode that suits you best in the settings. The sequence .+ means 'one or more occurrences of an arbitrary character' in the regular expressions.

If you are going to study affixes or other grapheme sequences that usually only occur in a certain word class, we suggest that, as the first step, you perform your planned character string search without restricting it to a certain word class. In this way, you can check your hypothesis about the distribution of the searched grapheme sequence in different word classes. After taking this precaution, you can go on to do searches that are restricted to certain word classes. The reason for this advice is that there may be differences and inconsistencies in the way in which world classes are coded in different languages.

The user should remember that the word lists contain both literary and dialectal vocabulary. Words belonging to the same word family can sometimes be found in two or more places if some component of the lexeme appears in both literary and dialectal form.

7. Acknowledgements

The present package containing three word lists and the user interface program is the fruit of coordinated collaboration between researchers and programmers from five universities. The compilers of the word lists wish to express their gratitude to the following people (in alphabetical order): Andrei Boltachev, T. I. Ibrahimov, F. M. Khisamova, Arto Moisio, Svetlana Saadat and Sirkka Saarinen. The names of the developers of the user interface program's earlier version 1.3 can be found through the program's Help menu, see "About SFOu WordListTool…". The present version 1.4 was developed by Dawid Adamkiewicz and Tomasz Ścibiorek from Poland. The Finnish institutions that have had the most central role in the realisation of this publication are the Research Unit for Volgaic Languages at the University of Turku, the Degree Programme in Information Technology at Turku University of Applied Sciences, the Centre for International Mobility (Helsinki) and the Finno-Ugrian Society (Helsinki). The work on the Tatar word list was supported by the Kone Foundation.

8. References

Enye Lav & Luutonen, J. 2012: *Reverse Dictionary of Komi (Zyryan). Обратный словарь коми (зырянского) языка.* Lexica Societatis Fenno-Ugricae XXXIV. Helsinki: Société Finno-Ougrienne.

Luutonen, J. et al. (ed.) 2007: *Electronic Word Lists: Mari, Mordvin and Udmurt. With SFOu WordListTool 1.3.* Lexica Societatis Fenno-Ugricae XXXI:1. Helsinki: Société Finno-Ougrienne.

Luutonen, J. & Mosin, M. & Shchankina, V. 2004: *Reverse Dictionary of Mordvin. Обратный словарь мордовских языков.* Lexica Societatis Fenno-Ugricae XXIX. Helsinki: Société Finno-Ougrienne.

Luutonen, J. & Saarinen, S. & Moisio, A. & Sergeyev, O. & Matrosova, L. 2002: *Reverse Dictionary of Mari (Cheremis). Обратный словарь марийского языка.* Lexica Societatis Fenno-Ugricae XXVIII. Helsinki: Société Finno-Ougrienne.

Насибуллин, Р. Ш. & Дудоров, В. Ю. 1992: *Обратный словарь удмуртского языка.* Ижевск: Удмуртский университет.

Zheltov, P. & Fomin, E. & Luutonen, J. 2009: *Reverse Dictionary of Chuvash. Обратный словарь чувашского языка.* Lexica Societatis Fenno-Ugricae XXXIII. Helsinki: Société Finno-Ougrienne.

Appendix:
Description of some files in the *alphabets*, *data* and *documentation* folders

The *alphabets* subfolder of the SFOu WordListTool program folder contains the following files, given here with their roles in the program:

Alph_order_Lat&Cyr.txt defines the alphabetical order;

characters.txt defines the buttons of the virtual keyboard;

defIgnored.txt defines the characters that are ignored when the lists are alphabetised.

The *data* subfolder of the program folder:

languages.txt defines the texts in tabs, buttons, dialogs, etc. of the user interface in the three languages.

The user can edit all of the abovementioned files to customise the *SFOu WordListTool*. For instructions, see *User_gude_en.pdf*.

The *documentation* subfolder of the SFOu WordListTool program folder contains the following files:

QuickStartManual_en.pdf (the document you are reading);

QuickStartManual_ru.pdf (the Quick Start Manual in Russian);

QuickStartManual_fi.pdf (the Quick Start Manual in Finnish);

Descriptions_en_2016.pdf (a detailed description of the word lists in English);

Descriptions_ru_2016.pdf (a detailed description of the word lists in Russian);

Tat_sanalista_kuvaus.pdf (a detailed description of the Tatar word list in Finnish);

Chu_source_h.pdf (Russian translations for the words with the source abbreviation *h* in the Chuvash word list);

User_guide_en.pdf (the user guide for the *SFOu WordListTool* in English);

User_guide_ru.pdf (the user guide for the *SFOu WordListTool* in Russian);

User_guide_fi.pdf (the user guide for the *SFOu WordListTool* in Finnish);

WordLists_licence_2016.pdf (the licence text for the lists in English and Finnish);

SFOuWLT_licence_2016.pdf (the licence text for the program in English and Finnish).