

CORPUS OF WRITTEN TATAR

User's Guide

v3.4

Mansur Saykhunov

Tavzikh Ibragimov

Jorma Luutonen

Contents

[What is the Corpus of Written Tatar?](#)

[Why is the language corpus necessary?](#)

[How will the Corpus of Written Tatar be developed?](#)

[System requirements](#)

[How to do a search in The Corpus of Written Tatar?](#)

[How to get back to the search page?](#)

[How to get a direct link to a specific page of search results?](#)

[What to do if you don't have a Tatar keyboard layout?](#)

[How to find a word form and reveal its contextual statistics in the Corpus?](#)

[How to find a lexeme in the Corpus?](#)

[How to find all lexemes beginning with a particular prefix in the Corpus?](#)

[How to find all lexemes ending in a given postfix in the Corpus?](#)

[How to find a certain word combination in the Corpus?](#)

[The Complex morphological search system:](#)

[About this system](#)

[How to find combinations of defined word forms?](#)

[How to find combinations using lemmas?](#)

[How to find combinations using morphological \(grammatical\) tags?](#)

[How to find combinations using the beginning, the middle and/or the end of words?](#)

[How to find combinations using several parameters for every word?](#)

[Entering search parameters in the graphical mode](#)

[How to listen to the sentences found in the Corpus?](#)

[Additional statistical materials!](#)

[Publications](#)

[Description](#)

[Use and citation](#)

[About the project](#)

[Acknowledgments](#)

[Contacts](#)

[In other languages](#)

What is the Corpus of Written Tatar?

The Corpus represents modern written Tatar language in electronic form. It consists of over 116 million words; the number of different word forms is 1.5 million. Materials publicly available on various WEB resources have been used in creating the Corpus. More than half of the texts ($\approx 60\%$ of the total volume of the Corpus) represents the publicistic style. Other styles covered by the Corpus are the following:

- fiction $\approx 35\%$,
- scientific literature (Humanities) $\approx 4\%$,
- official and business papers $\approx 1\%$.

The main sources of linguistic material were the most popular web sites, e.g. “Татар – информ» (Tatar – inform), “Ислам – татарлар һәм мөселманнар (Islam – Tatars and muslims)”, “Татар электрон китапханәсе (Tatar digital library)”, “Дуслык (Friendship)”, “Азатлык” (Freedom), “Яңа гасыр (New century)”, “Татар әдипләре (Tatar writers)”, etc.

The Corpus of Written Tatar is intended for research on the lexico-semantic system of the Tatar language in the framework of statistical lexicology and cognitive linguistics.

The Corpus can be used in three modes:

- the statistical (search for a given word, determining its left and right neighbours, calculating the frequency of use of word forms);
- lemmatization (due to the presence of morphological marking it is possible to look for examples with all grammatical forms of a given lemma);
- pattern matching search (by initial, middle or final part of the word).

Furthermore, it is possible to listen to the sentences shown as examples of the use of the given word. If necessary, the corpus manager can extend the right and left contexts of a given word and thereby increase the accuracy in determining the meaning of a given word or phrase.

The search engine of the Corpus of Tatar language allows to perform the following operations:

- to conduct search for given words, to reveal the frequency of their use, to find examples confirming the use of given word or word form in language;
- to determine which words can occur in front of, or after the word (left and right contexts of a given word), and frequency rates of context-related words in combination with this word.

Why is the language corpus necessary?

The composition of the corpus, not making any strict divisions between functional styles but offering the user a large volume of material, adequately reflects the current state of the Tatar language. A large number of texts written by the most able people not only reveals the world-outlook of authors, but also that of the readers, and, further, the world-view of the whole ethnic community of the epoch in question. The Corpus of Written Tatar can thus be viewed as a monument of the stage of the society's historical development in that period.

How will the Corpus of Written Tatar be developed?

One of the most important linguists of the 20th century, L. Bloomfield finishes his book "Language" with the words "It is only a prospect, but not hopelessly remote, that the study of language may help us toward the understanding and control of human events". Realization of the Bloomfieldian project requires, among other things, creation of different types of subcorpora, electronic dictionaries - both usual ones and thesauruses.

There will be a need of interpretation of how the language community accumulates life experience, learns and transforms the world where it lives. It is obvious that the specified requirements will also define the ways of corpus linguistics development.

System requirements

For problem-free work with The Corpus, we recommend:

1. Use the most recent versions of web-browsers:

- Mozilla Firefox: <https://www.mozilla.org/en-US/firefox/all/>
- Google Chrome: <https://www.google.ru/intl/en/chrome/browser/desktop/index.html>

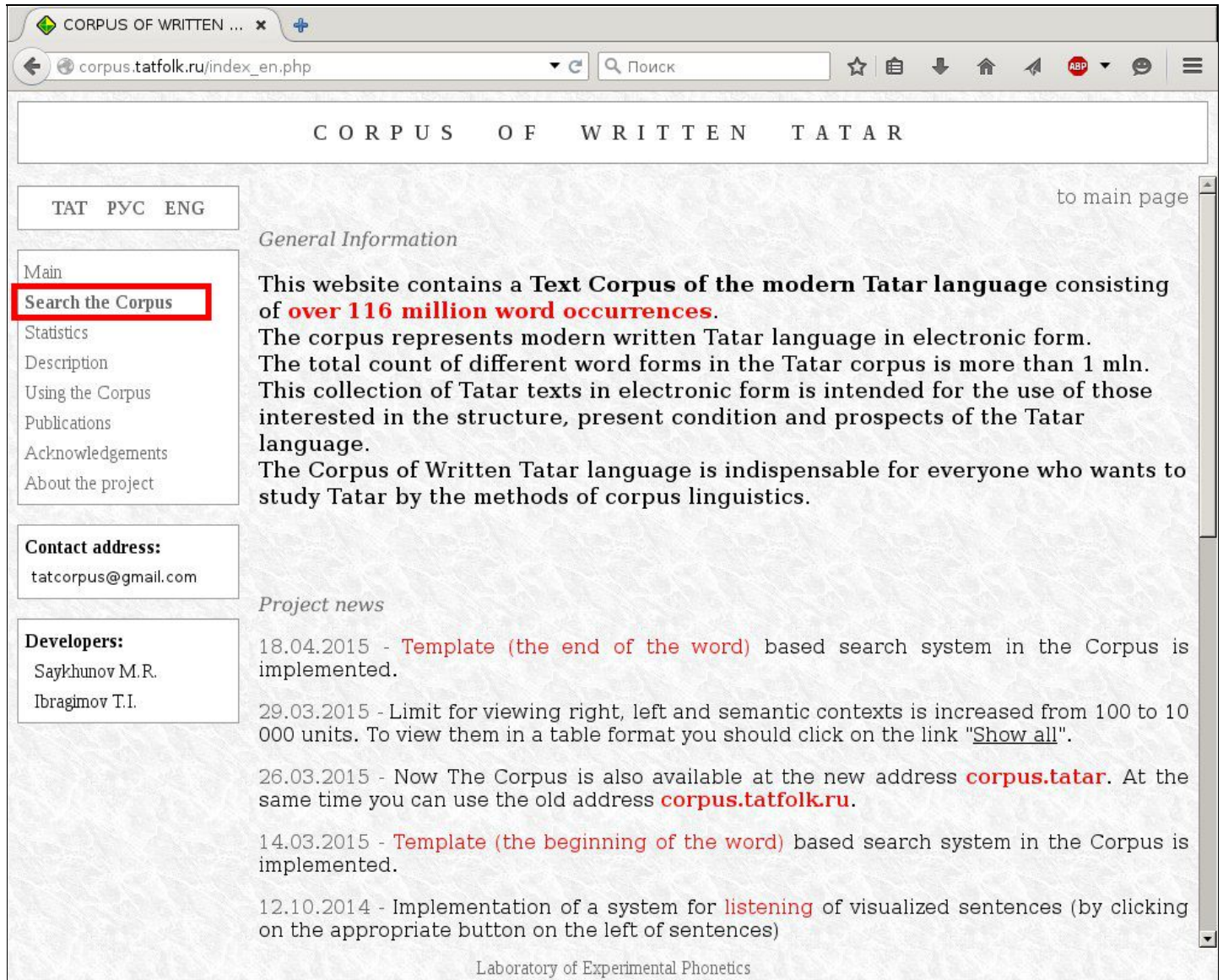
2. Turn on JavaScript support in the web-browser.

3. If you have problems with listening to the sentences:

- Try to use another web-browser
- Install new version of Adobe Flash Player (<https://get.adobe.com/flashplayer/>)

How to do a search in The Corpus of Written Tatar?

On the home page of corpus.tatar or corpus.tatfolk.ru web-site, find the link "Search the Corpus" in the left menu and click it:



The screenshot shows a web browser window with the address bar displaying corpus.tatfolk.ru/index_en.php. The page title is "CORPUS OF WRITTEN TATAR". In the top left corner, there are language selection buttons: "TAT PYC ENG". A navigation menu on the left side contains the following items: "Main", "Search the Corpus" (highlighted with a red box), "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are sections for "Contact address:" (tatcorpus@gmail.com) and "Developers:" (Saykhunov M.R., Ibragimov T.I.). The main content area features a "General Information" section with the following text: "This website contains a **Text Corpus of the modern Tatar language** consisting of **over 116 million word occurrences**. The corpus represents modern written Tatar language in electronic form. The total count of different word forms in the Tatar corpus is more than 1 mln. This collection of Tatar texts in electronic form is intended for the use of those interested in the structure, present condition and prospects of the Tatar language. The Corpus of Written Tatar language is indispensable for everyone who wants to study Tatar by the methods of corpus linguistics." Below this is a "Project news" section with several entries: "18.04.2015 - **Template (the end of the word)** based search system in the Corpus is implemented.", "29.03.2015 - Limit for viewing right, left and semantic contexts is increased from 100 to 10 000 units. To view them in a table format you should click on the link "[Show all](#)".", "26.03.2015 - Now The Corpus is also available at the new address **corpus.tatar**. At the same time you can use the old address **corpus.tatfolk.ru**.", "14.03.2015 - **Template (the beginning of the word)** based search system in the Corpus is implemented.", and "12.10.2014 - Implementation of a system for **listening** of visualized sentences (by clicking on the appropriate button on the left of sentences)". At the bottom of the page, it says "Laboratory of Experimental Phonetics".

As a result, the "Search in the Corpus of Tatar language" page will open:

CORPUS OF WRITTEN TATAR

TAT PYC ENG [to main page](#)

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдәгез**)

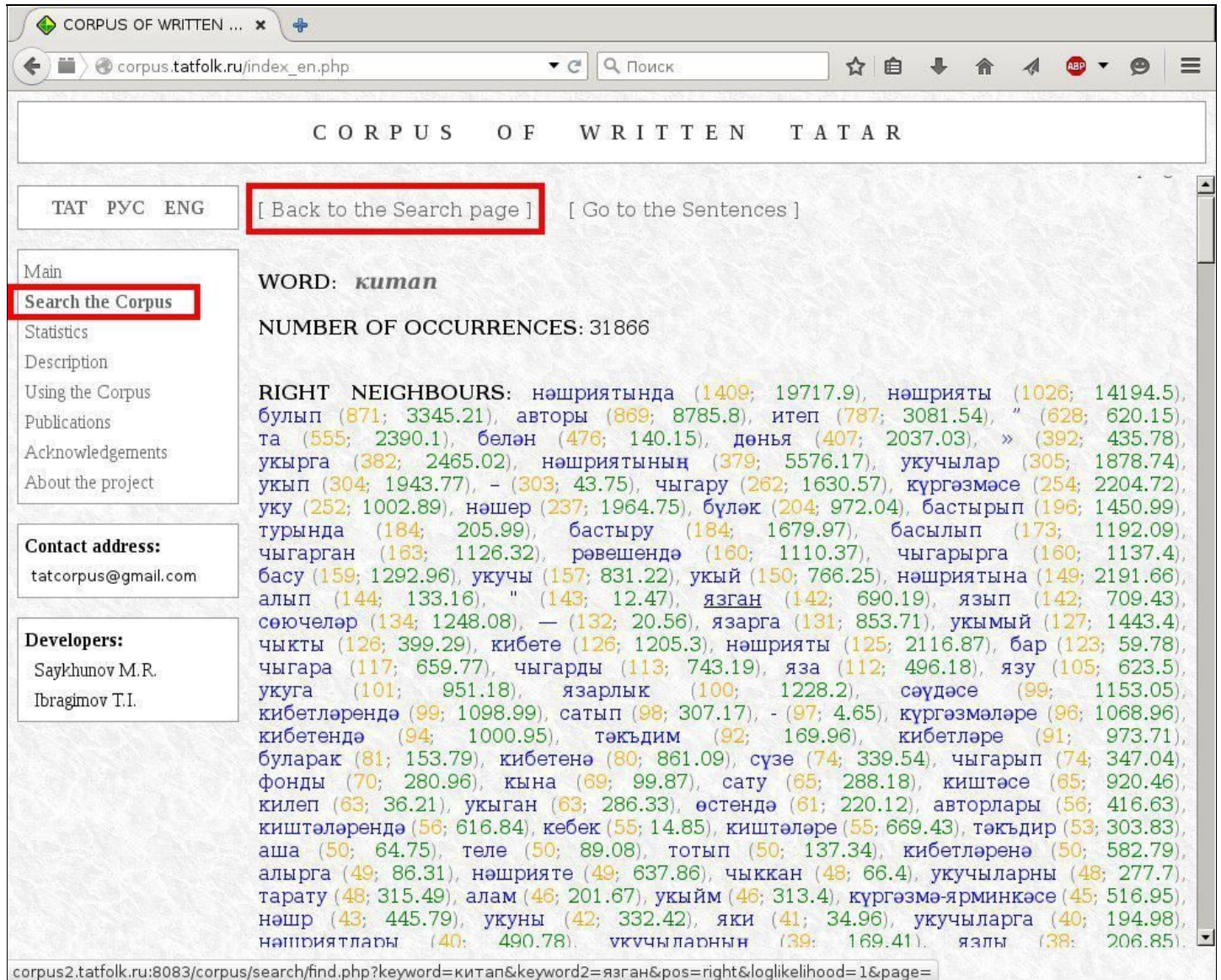
Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)

Laboratory of Experimental Phonetics

How to get back to the search page?

Having conducted a search query, you can return to the page of search parameters by using the following links:



The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". At the top, there are navigation links: "TAT РУС ENG", "[Back to the Search page]" (highlighted with a red box), and "[Go to the Sentences]". On the left sidebar, "Search the Corpus" is also highlighted with a red box. The main content area displays the search results for the word "kuman".

WORD: kuman
NUMBER OF OCCURRENCES: 31866

RIGHT NEIGHBOURS: нәшриятында (1409; 19717.9), нәшрияты (1026; 14194.5), булып (871; 3345.21), авторы (869; 8785.8), итеп (787; 3081.54), " (628; 620.15), та (555; 2390.1), белән (476; 140.15), дөнья (407; 2037.03), » (392; 435.78), укырга (382; 2465.02), нәшриятының (379; 5576.17), укучылар (305; 1878.74), укып (304; 1943.77), - (303; 43.75), чыгару (262; 1630.57), күргәзмәсе (254; 2204.72), уку (252; 1002.89), нәшер (237; 1964.75), бүләк (204; 972.04), бастырып (196; 1450.99), турында (184; 205.99), бастыру (184; 1679.97), басылып (173; 1192.09), чыгарган (163; 1126.32), рәвешендә (160; 1110.37), чыгарырга (160; 1137.4), басу (159; 1292.96), укучы (157; 831.22), укый (150; 766.25), нәшриятына (149; 2191.66), алып (144; 133.16), " (143; 12.47), язган (142; 690.19), язып (142; 709.43), сөючеләр (134; 1248.08), — (132; 20.56), язарга (131; 853.71), укымый (127; 1443.4), чыкты (126; 399.29), кибете (126; 1205.3), нәшрияты (125; 2116.87), бар (123; 59.78), чыгара (117; 659.77), чыгарды (113; 743.19), яза (112; 496.18), язу (105; 623.5), укуга (101; 951.18), язарлык (100; 1228.2), сәүдәсе (99; 1153.05), кибетләрендә (99; 1098.99), сатып (98; 307.17), - (97; 4.65), күргәзмәләре (96; 1068.96), кибетендә (94; 1000.95), тәкъдим (92; 169.96), кибетләре (91; 973.71), буларак (81; 153.79), кибетенә (80; 861.09), сүзе (74; 339.54), чыгарып (74; 347.04), фонды (70; 280.96), кына (69; 99.87), сату (65; 288.18), киштәсе (65; 920.46), килеп (63; 36.21), укыган (63; 286.33), өстендә (61; 220.12), авторлары (56; 416.63), киштәләрендә (56; 616.84), кебек (55; 14.85), киштәләре (55; 669.43), тәкъдир (53; 303.83), аша (50; 64.75), теле (50; 89.08), тотып (50; 137.34), кибетләренә (50; 582.79), алырга (49; 86.31), нәшрияте (49; 637.86), чыккан (48; 66.4), укучыларны (48; 277.7), тарату (48; 315.49), алам (46; 201.67), укыйм (46; 313.4), күргәзмә-ярминкәсе (45; 516.95), нәшр (43; 445.79), укуны (42; 332.42), яки (41; 34.96), укучыларга (40; 194.98), нәшриятлары (40; 490.78), укучыларның (39; 169.41), язлы (38; 206.85).

corpustatfolk.ru:8083/corpus/search/find.php?keyword=китан&keyword2=язган&pos=right&loglikelihood=1&page=

There is also a link at the bottom of the page:

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left side, there is a navigation menu with options: "TAT РУС ENG", "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are sections for "Contact address: tatcorpus@gmail.com" and "Developers: Saykhunov M.R., Ibragimov T.I.". The main content area displays search results for the word "kuman". Each result is preceded by a play button icon and includes a source citation and a "Find text" link. The results are:

- ▶ Мондый кызыклы чаралар **kuman** кибетендә һәр атнада үтөп торачак.
(source: "Белем.ру" порталы (web-caim)) [Find text](#)
- ▶ Шулай ук телевизордан яңгыраган куркыныч хәбәрләргә йөрәккә якин алмаска, үзгә берәр мавыктыргыч шөгьль табарга, кызыклы **kuman** укырга, күбрәк жыләк-жимеш һәм шоколад ашарга кушалар.
(source: "Татарстан яшьләре" газетасы (web-caim)) [Find text](#)
- ▶ Күрәсен, безнең халыкка **kuman** байлыгы түгел, ә башка байлык кыйммәтрәк, чөнки бик кыйммәт, шундый затлы китаплар ун сум, шуның кадәр төшерделәр **kuman** бәясен, китапның бәясе төшәргә тиеш түгел, чөнки ул бик күп көшенә хезмәт.
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caim)) [Find text](#)
- ▶ **Kuman** бизәү өлкәсендә «Жайдак» үзен тиз таныта.
(source: Максим Глухов. Эзләнүләр жимеше)
- ▶ Башка кунакларны көткөндә, **kuman** сүзләрен сөйләләр иде.
(source: "Азатлык Радиосы" (web-caim)) [Find text](#)
- ▶ 1) Дүрт йөзгә якин фотосурәт төшергән фотографларның яшен һәм елъязманның Калининградта чыгарылуын да исәпкә алсаң, **kuman** шактый гына бәяләгә охшап тора.
(source: Гульшат Галиуллина. Татар теле: лексикология (таблицалар, анализ үрнәкләре, күнегүләр, сүзлекчә). КАЗАН - 2007)

At the bottom of the page, there are navigation links: "< Prev | Begin | Next >". Below these, there are two buttons: "[Top]" and "[Back to the Search page]", where the latter is highlighted with a red rectangular box. The footer of the page reads "Laboratory of Experimental Phonetics".

How to get a direct link to a specific page of search results?

Please do not try to reference the search results with the url in the browser's address bar. Since the Corpus site has a frame-based structure, url is just a link to the home page. To get an appropriate link, you can use the "Direct Link" button at the top of the search results page:

The screenshot shows the search results page for the word "kuman" on the Corpus of Written Tatar website. The page is titled "CORPUS OF WRITTEN TATAR" and includes navigation links for "TAT", "РУС", and "ENG". A red box highlights the "[Direct link]" button. The search results show the word "kuman" with a play button icon and a count of 31866 occurrences. The "RIGHT NEIGHBOURS" section lists various words and their frequencies, such as "нәшриятында" (1409; 19717.9) and "авторы" (869; 8785.8). The "LEFT NEIGHBOURS" section lists words like "%^%" (2870; 265.87) and "татарстан" (2781; 16768.4). The page footer includes the text "Laboratory of Experimental Phonetics".

CORPUS OF WRITTEN TATAR

TAT РУС ENG [Back to the Search page] [Go to the Sentences] [Direct link] to main page

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project
Useful information

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

WORD: *kuman*

NUMBER OF OCCURRENCES: 31866

RIGHT NEIGHBOURS: нәшриятында (1409; 19717.9), нәшрияты (1026; 14194.5), булып (871; 3345.21), авторы (869; 8785.8), итеп (787; 3081.54), " (628; 620.15), та (555; 2390.1), белән (476; 140.15), дөнья (407; 2037.03), » (392; 435.78), укырга (382; 2465.02), нәшриятының (379; 5576.17), укучылар (305; 1878.74), укып (304; 1943.77), - (303; 43.75), чыгару (262; 1630.57), күргәзмәсе (254; 2204.72), уку (252; 1002.89), нәшер (237; 1964.75), бүлөк (204; 972.04), бастырып (196; 1450.99), турында (184; 205.99), бастыру (184; 1679.97), басылып (173; 1192.09), чыгарган (163; 1126.32), рәвешендә (160; 1110.37), чыгарырга (160; 1137.4), басу (159; 1292.96), укучы (157; 831.22), укый (150; 766.25), нәшриятына (149; 2191.66), алып (144; 133.16), " (143; 12.47), язган (142; 690.19), язып (142; 709.43), сөжүчләр (134; 1248.08), — (132; 20.56), язарга (131; 853.71), укымый (127; 1443.4), чыкты (126; 399.29), кибете (126; 1205.3), нәшрияты (125; 2116.87), бар (123; 59.78), чыгара (117; 659.77), чыгарды (113; 743.19), яза (112; 496.18), язу (105; 623.5), укуга (101; 951.18), язарлык (100; 1228.2), сөздәсе (99; 1153.05), кибетләрендә (99; 1098.99), сатып (98; 307.17), - (97; 4.65), күргәзмәләре (96; 1068.96), кибетендә (94; 1000.95), тәкъдим (92; 169.96), кибетләре (91; 973.71), буларак (81; 153.79), кибетенә (80; 861.09), сүзә (74; 339.54), чыгарып (74; 347.04), фонды (70; 280.96), кына (69; 99.87), сату (65; 288.18), киштәсе (65; 920.46), килеп (63; 36.21), укыган (63; 286.33), өстендә (61; 220.12), авторлары (56; 416.63), киштәләрендә (56; 616.84), көбек (55; 14.85), киштәләре (55; 669.43), тәкъдир (53; 303.83), аша (50; 64.75), теле (50; 89.08), тотып (50; 137.34), кибетләренә (50; 582.79), алырга (49; 86.31), нәшрияте (49; 637.86), чыккан (48; 66.4), укучыларны (48; 277.7), тарату (48; 315.49), алам (46; 201.67), укыйм (46; 313.4), күргәзмә-ярминкәсе (45; 516.95), нәшр (43; 445.79), укуны (42; 332.42), яки (41; 34.96), укучыларга (40; 194.98), нәшриятлары (40; 490.78), укучыларның (39; 169.41), язды (38; 206.85), дөньясына (38; 214.69), нәшриятендә (38; 487.07), сөжүчә (38; 309.94), битләре (38; 318.18), киштәсендә (37; 467.97), битләренә (36; 322.77), бастыра (36; 229.48), палатасы (35; 151.32) [Show all](#)


LEFT NEIGHBOURS: %^% (2870; 265.87), татарстан (2781; 16768.4), , (2522; 144.94), бу (1203; 2840.53), аерым (771; 4948.61), " (709; 742.16), бер (680; 1378.6), _NUMBER_ (660; 66.09), дигән (659; 2501.77), турында (502; 1398.4), « (365; 325.43), алегә (362; 926.18), яна (359; 865.88), исемле (350; 2094.36)

Laboratory of Experimental Phonetics

As a result, a special window will open, showing you the desired link, which you now can copy:

CORPUS OF WRITTEN TATAR

TAT РУС ENG [Back to the Search page] [Go to the Sentences] [Direct link] to main page

WORD: *kuman* 

NUMBER OF OCCURRENCES: 31866

The page at <http://corpus2.tatfolk.ru:8083> says:

Direct link

corpus/search/find.php?keyword=%D0%BA%D0%B8%D1%82%D0%BA

RIGHT NEIGHBOURS: авторы (869; 14194.5), дөнья (407; 2390.1), күргәзмәсе (306; 5576.17), бастырып (100; 1630.57), чыгарган (100; 1964.75), укучы (157; 972.04), язган (142; 690.19), язып (142; 709.43), сөүчөләр (134; 1248.08), — (132; 20.56), язарга (131; 853.71), укумый (127; 1443.4), чыкты (126; 399.29), кибете (126; 1205.3), нәшрияты (125; 2116.87), бар (123; 59.78), чыгара (117; 659.77), чыгарды (113; 743.19), яза (112; 496.18), язу (105; 623.5), укуга (101; 951.18), язарлык (100; 1228.2), сөүдәсе (99; 1153.05), кибетләрендә (99; 1098.99), сатып (98; 307.17), - (97; 4.65), күргәзмәләре (96; 1068.96), кибетендә (94; 1000.95), төкъдим (92; 169.96), кибетләре (91; 973.71), буларак (81; 153.79), кибетенә (80; 861.09), сүзе (74; 339.54), чыгарып (74; 347.04), фонды (70; 280.96), кына (69; 99.87), сату (65; 288.18), киштәсе (65; 920.46), килеп (63; 36.21), укыган (63; 286.33), өстендә (61; 220.12), авторлары (56; 416.63), киштәләрендә (56; 616.84), кебек (55; 14.85), киштәләре (55; 669.43), төкъдир (53; 303.83), аша (50; 64.75), төле (50; 89.08), тотып (50; 137.34), кибетләренә (50; 582.79), алырга (49; 86.31), нәшрияте (49; 637.86), чыккан (48; 66.4), укучыларны (48; 277.7), тарату (48; 315.49), алам (46; 201.67), укыйм (46; 313.4), күргәзмә-ярминкәсе (45; 516.95), нәшр (43; 445.79), укуны (42; 332.42), яки (41; 34.96), укучыларга (40; 194.98), нәшриятлары (40; 490.78), укучыларның (39; 169.41), язды (38; 206.85), дөньясына (38; 214.69), нәшриятендә (38; 487.07), сөүчө (38; 309.94), битләре (38; 318.18), киштәсендә (37; 467.97), битләренә (36; 322.77), бастыра (36; 229.48), палатасы (35; 151.32) [Show all](#)

LEFT NEIGHBOURS: %^% (2870; 265.87), татарстан (2781; 16768.4), , (2522; 144.94), бу (1203; 2840.53), аерым (771; 4948.61), " (709; 742.16), бер (680; 1378.6), NUMBER (660; 66.09), дигән (659; 2501.77), түрүндә (502; 1398.4), « (365; 325.43), әлеге (362; 926.18), яна (359; 865.88), исемле (350; 2094.36)

Laboratory of Experimental Phonetics

What to do if you don't have a Tatar keyboard layout?

If no Tatar keyboard has been installed (or configured) into your computer, you can use the virtual keyboard. You find the virtual keyboard icon on the right side of the text field in the search page. Using this facility, you can type the desired word or word form directly from the screen. After typing with the virtual keyboard, you normally click the "**Find!**" button to launch the search.

CORPUS OF WRITTEN TATAR

TAT РУС ENG [to main page](#)


Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Search the Corpus of Tatar language

Select the search type and enter the desired word:

 Find!

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмөдегез**)

Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *лөргө**)

Laboratory of Experimental Phonetics

Using the virtual keyboard is very easy. When it pops up, you just write the word you want to find by clicking the letters in the virtual keyboard, or, alternatively, by pressing the corresponding buttons on your own physical keyboard.

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left, there is a navigation menu with links: "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are sections for "Contact address: tatcorpus@gmail.com" and "Developers: Saykhunov M.R., Ibragimov T.I.". The main content area features a search interface with the heading "Search the Corpus of Tatar language" and the instruction "Select the search type and enter the desired word:". There are three radio buttons for search options: "Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)", "Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)", and "Pattern character". A search input field and a "Find!" button are present. A virtual keyboard window titled "Virtual Keyboard - Mozilla Firefox" is overlaid on the page, showing a Tatar keyboard layout with keys for letters like "й, ц, е, у, к, е, н, г, ш, щ, ә, з, х, ь, ү" and "ф, ы, в, а, п, р, о, л, д, ж, ң, э". The keyboard also includes standard function keys like Tab, CapsLock, Shift, Ctrl, and Alt. The version "VirtualKeyboard 3.7.2.792" and a language dropdown set to "Tatar" are visible at the bottom of the keyboard window.

How to find a word form and reveal its contextual statistics in the Corpus?

Type the word or word form you want to examine (e.g., *китапның*) in the text field on the search page; make sure that the «**Search for collocations in the contextual (statistic) corpus by wordform...**» checkbox below is selected and press the «**Find!**» button:

CORPUS OF WRITTEN TATAR

TAT РУС ENG [to main page](#)

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Search the Corpus of Tatar language

Select the search type and enter the desired word:

китапның Find!

Search for collocations in the contextual (statistic) corpus by wordform (for example, китапны, авылларга, килмәдергез)

Search in the morphologically annotated corpus for lemma (for example, китап, авыл, кил)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, китап*, авыл*, *гез, *ләргә)

Laboratory of Experimental Phonetics

Search results will open in a new window. First, the targeted word and the number of its occurrences in the Corpus will be shown. Below these, further data is given in three sections:

- **Right neighbours** - a list of words, sorted by frequency in descending order, that appear in positions to the right of the targeted word in the Corpus. That means those words that are located after the searched word in the sentences where this word is found. Next to the word, two numbers are given in parentheses:
 1. The number highlighted in **orange** color indicates the amount of occurrences of this combination of words in the Corpus, e.g., the targeted *китапның* + current word from the right context *эчтәлеген*. In this case, it turns out that the collocation «*китапның эчтәлеген*» was found 26 times in the Corpus.
 2. The number highlighted in **green** represents the **log-likelihood ratio** for the current collocation in the Corpus. More information about this concept can be found in the personal blog of Ted Dunning at: <http://tdunning.blogspot.ru/2008/03/surprise-and-coincidence.html>

CORPUS OF WRITTEN TATAR

TAT РУС ENG WORD: *китапның*
NUMBER OF OCCURRENCES: 3120

RIGHT NEIGHBOURS: авторы (146; 1625.81), _NUMBER_ (144; 134.06), беренче (119; 579.32), икенче (86; 475.46), исеме (73; 567.74), төп (56; 282.71), исем (55; 453.97), бер (52; 84.03), соңгы (50; 232.22), эчтәлегә (44; 476.6), ахырында (42; 342.94), « (39; 39.47), һәр (39; 152.04), кереш (37; 400.51), “ (30; 3.95), иң (29; 59.37), авторлары (28; 299.1), тиражы (26; 323.12), **эчтәлеген (26; 285.34)**, мөхәррире (26; 226.15), исемен (22; 139.32), эчтәлегенә (21; 262.48), төзүчесе (20; 259.23), баш (17; 47.7), кадрен (16; 134.95), титул (16; 222.34), фәнни (15; 71.71), язылу (14; 117.73), әһәмияте (14; 113.48), дөньяга (13; 62.35), тышлыгы (13; 190.91), тышлыгында (13; 192.15), яңа (13; 10.95), тагын (13; 19.43), электрон (12; 62.8), кыйммәте (12; 122.25), редакторы (12; 109.51), ахырына (12; 77.81), рус (12; 40.54), туыландырылган (11; 145.04), басылып (11; 66.24), үз (11; 4.91), эчәнә (11; 58.35), башында (11; 48.32), дөнья (11; 28.38), нинди (11; 21.16), берничә (11; 26.21), зурлыгы (11; 110.07), дәвам (10; 89.46), азагында (10; 69.35), ике (10; 7.43), эчәндә (9; 23.93), аерым (9; 21.22), беренчесе (9; 61.76), баясе (9; 45.17), күп (9; 4.68), тарихы (9; 38.27), үзен (9; 30.2), тышлыгын (9; 143.19), исемәнә (9; 52.86), өченчә (9; 31.36), дәрәжәсен (8; 48.51), рәссамы (8; 66.54), үзенә (8; 23.84), чыгуы (8; 57.27), тышлыгына (7; 98.96), идея (7; 47.72), буеннан-буена (7; 78.26), илаһи (7; 49.64), тулы (7; 19.42), тышына (7; 80.96), максаты (7; 28.99), кулъязмасы (7; 74.51), тәржемәсен (7; 82.86), концепциясе (7; 61.05), дүртөнчә (7; 35.75), ролен (6; 34.36), алгы (6; 40.17), берсен (6; 33.5), тәүге (6; 28.69), битләрен (6; 57.76), баясен (6; 41.32), нигезәнә (6; 48.04), урыны (6; 21.71), җаваплы (6; 23.14), нигезен (6; 43.45), авырлыгы (6; 44.28), барлык (6; 7.04), калган (6; 9.22), киңәйтәлгән (5; 36.03), битләре (5; 44.65), тәржемәсе (5; 42.59), архив (5; 34.78), озаклай (5; 31.25), тышында (5; 56.4), баштагы (5; 41.35), кырыена (5; 45.43), россиягә (5; 34.23), эчтәлегендә (5; 61.1), рецензенты (5; 86) [Show all](#)

L... NEIGHBOURS ...

Laboratory of Experimental Phonetics

- **Left neighbours** - the same as in the preceding paragraph but applied to the left-side context of the word, i.e. what word is found immediately before the targeted word!
- **Co-occurrences inside the sentence** - a list of words, sorted by frequency in descending order, that appear in the sentences where the targeted word is found. The co-occurring word may be located next to the targeted word, or be separated from it by other words. The co-occurrences reflect various semantic relations between the words in the sentence.

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left side, there is a navigation menu with options: "TAT", "РУС", "ENG", "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", "About the project", "Contact address: tatcorpus@gmail.com", and "Developers: Saykhunov M.R., Ibragimov T.I.". The main content area displays search results for "LEFT NEIGHBOURS" and "CO-OCCURRENCES".

LEFT NEIGHBOURS: %^% (903; 1453.07), бу (305; 1270.4), , (224; 4.69), изге (150; 1333.51), әлеге (119; 576.55), ул (69; 103.51), дигән (47; 150.31), исемле (30; 171.36), яңа (27; 52.82), һәр (24; 71.81), шул (24; 33.4), бер (22; 9.61), чыккан (19; 72.8), ук (18; 19.23), булачак (17; 52.69), турындагы (16; 65.46), ә (16; 8.43), аталган (12; 62.09), шушы (12; 21.82), беренче (12; 10.21), икенче (12; 22.34), торган (11; 8.26), әһле (11; 128.83), берничә (11; 26.21), әмма (11; 15.16), язылган (10; 37.13), кулъязма (10; 93.08), күргән (9; 39.25), багышланган (9; 26.72), автор (9; 46.24), басма (9; 60.61), өч (9; 17.26), борынгы (8; 31.92), ителгән (7; 16.58), мондый (7; 9.98), тик (7; 7.75), чөнки (7; 8.42), миңа (7; 12.19), исемнәре (6; 33.04), басылган (6; 32.22), тарихи (6; 17.88), алган (6; 7.74), алынган (6; 14.72), кадрле (6; 27.01), шигырьләр (5; 20.66), торучы (5; 20.74), томлы (5; 51.84), кызыл (5; 13.67), киләчәктә (5; 13.1), текст (5; 34.75), томлык (4; 34.99), биографик-публицистик (4; 86.32), исемләнгән (4; 17.56), фәнни (4; 9.3), хәтта (4; 4.64), исеме (4; 8.84), андый (4; 9.43), кызыксыну (3; 11.72), кыскача (3; 16), битлек (3; 22.31), бүленгән (3; 15.88), тормышында (3; 12.86), мөшһүр (3; 12.91), тышлы (3; 27.7), уникаль (3; 15.58), калын (3; 15.57), ошбу (3; 16.82), данә (3; 16.46), ушбу (3; 29.99), аңлатмалар (3; 24.08), электрон (3; 7.87), күнелле (3; 6.62), яза (3; 6.16), * (3; 5.43), әлеге (3; 4.41), жиде (3; 8.44), һәрбер (3; 8.87), теләгән (3; 9.18), гадәти (3; 10.6), моны (3; 4.01), әсәрләрдә (3; 23.57), ләп (3; 21.02), чыгарылган (3; 11.69), кирәкле (3; 7.54), битле (2; 14.27), мөхәммәтшиннар (2; 27.08), шәжәрәсе (2; 16.98), кәшифәгә (2; 34.78), битенә (2; 12.49), жәмәгатьчелегәндә (2; 20.1), исемдәге (2; 12.57), көйле (2; 15.02), авторлар (2; 11.45), габидулла (2; 24.94), жавапны (2; 11.57), хикмәтле (2; 13), бөлкәм (2; 10.89), бар—алары (2; 43.16), әзерләнүче (2; 17.86), фикирләргә (2; 12.84) [Show all](#)

CO-OCCURRENCES: . (2934; 244.58), , (1741; 50.28), һәм (713; 187.99), _NUMBER (586; 39.54), бу (536; 338.75), да (431; 23.99), белән (417; 9.5), « (378; 334.19), » (353; 313.95), - (336; 24.63), дә (330; 17.62), " (318; 24.2), " (294; 22.12), ул (294; 24.14), дип (281; 39.63), : (260; 73.1), бер (254; 65.54), ((228; 142.83),) (223; 128.39), беренче (215; 249.58), турында (209; 179.09), татар (201; 116.14), авторы (201; 1294.23), әлеге (190; 179.29), дигән (187; 193.42), аның (171; 33.32), изге (164; 623.39), - (149; 19.47), китап (145; 527.08), шул (144; 39.24), ук (137; 39.36), икенче (137; 200.95), ә (136; 9.3)

Laboratory of Experimental Phonetics

Below the co-occurrences data, examples are given:

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left side, there is a navigation menu with options: "TAT РУС ENG", "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are sections for "Contact address: tatcorpus@gmail.com" and "Developers: Saykhunov M.R., Ibragimov T.I.". The main content area is titled "EXAMPLES:" and displays five search results for the word "китапның" (highlighted in red). Each result includes a play button icon, a text snippet, and a source citation with a "Find text" link.

EXAMPLES:

- ▶ Әлеге уңайдан Бөтендөнъя татар конгрессы Башкарма комитеты рәисе урынбасары Ренат Вәлиуллин **китапның** тиздән сатуга чыгачагын, шулай ук республикадан читтә дә таратылачагын белдерде.
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caim)) [Find text](#)
- ▶ Шуңа күрә дә китабымның изге максаты - жыйналган мәгълүматларны, тарихи чыганаclar белән тулыландырып, халкымның үзенә кайтару». 3 мең данә тираж белән нәшер ителгән **китапның**, тышлыгын рәссам Рәисә Сәйфуллина бизәгән.
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caim)) [Find text](#)
- ▶ Дүрт тематик кисәккә бүленгән **китапның** эпиграфлары ук игътибарны туплай, кызыксынуны көчәйтә.
(source: РИФӘ РАХМАН. Татар телендә 100 сочинение)
- ▶ **Китапның** икенче басмасы алдагы басмада жиберелгән күп кенә житешсезлекләрдән, авыр сүзләрдән азат ителгән.
(source: "Фән һәм тел" журналы (№27))
- ▶ Профессор Г. Ф. Саттаров "Казанчы, Сабанчы, Уракчы, Урманчы, Ындырчы, Буранчы, Туйчы, Ямгурчы һ. б. — борынгы төрки татар кешесе исемнәре" дип яза алда телгә алынган хезмәтендә (85нче бит.) "Башкортстанда феодализм һәм капитализм тарихыннан" дигән **китапның** (Уфа, 1971 ел) 267нче битендә "Слобоцкие татаровя без хлебного жалованья Иткул Бехтемиров, Сабанча Чилпанов..." дигән юлларны укыйбыз.
(source: "Кызыл таң" газетасы (web-caim)) [Find text](#)
- ▶ Хәлбуки, табыг иттерү ноктасыннан **китапның** озын булуы матлуб

Laboratory of Experimental Phonetics

About 50 sentences where the searched word is used are shown. For convenience, the word is highlighted in red.

The text in small print after each sentence indicates the source from which the example was taken.

If the source is a web resource ("**web-site**"), the link "**Find text**" will be available.

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left, there is a navigation menu with options: "TAT PYC ENG", "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there is a "Contact address:" section with the email `tatcorpus@gmail.com` and a "Developers:" section listing "Saykhunov M.R." and "Ibragimov T.I.". The main content area displays several search results for the word "китанның". Each result consists of a text snippet, a source attribution in small print, and a "Find text" link. The second result is highlighted with a red border. The footer of the page reads "Laboratory of Experimental Phonetics".

C O R P U S O F W R I T T E N T A T A R

TAT PYC ENG

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Хэлбуки, табыг иттерү ноктасыннан **китанның** озын булуы матлуб түгелдер.
(source: "Эхо веков" журналы)

Изге **Китанның** чираттагы басмасы 1694 елда Гамбургта (Германия) нашир ителә.
(source: "Ислам - Татарлар һәм мөселманнар" (web-caŭm)) [Find text](#)

Өйгә эш катлаулашты: әкиятне укып, үзеңә ошаган фрагментын сүрәткә төшерергә, һәм **китанның** ни хакында булуын бер жөмлә белән генә язып куярга кирәк.
(source: "Гыймрановларның сайты" (web-caŭm)) [Find text](#)

Китанның әһәмиятен якташыбыз, тарих фәннәре кандидаты, Казан федераль (педагогика) университеты профессоры, ТР Мәгълүматлаштыру академиясенең действительный әгъзасы Иршат Гафаров һәм чуваш төбәкләрен өйрәнү союзының мактау-лы рәисе Виталий Станьял югары бәяләделәр.
(source: "Туган як" газетасы (web-caŭm)) [Find text](#)

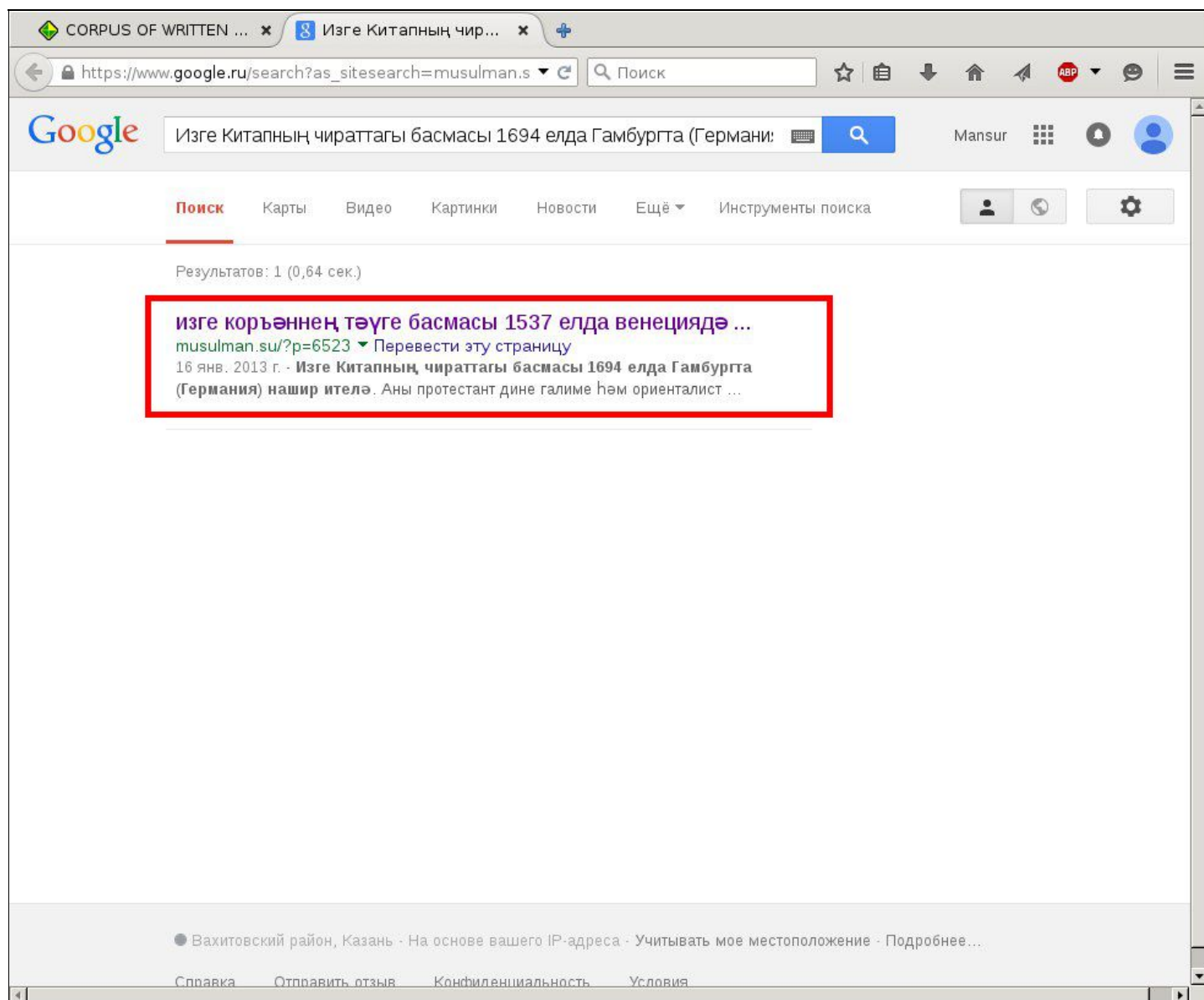
"Сагыну" дигән **китанның** кыйбласы менә шул тарафта.
(source: "Intertat.ru" электрон газетасы (web-caŭm)) [Find text](#)

Күпме хезмәт, чыгым тотып әзерләнгән әлеге **китанның** тузан жыеп ятуы шагыйрәне генә түгел, каләмдәшләрен дә борчуга салган һәм алар ярдәм сорап, УР Милли сәясәт министрлыгының милли эшләр буенча баш белгече Айрат Гайнетдиновка мөрәжәгать ителәр.
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caŭm)) [Find text](#)

Laboratory of Experimental Phonetics

In the case of belletristic texts, there is no "**Find text**" link because of restrictions related to copyright.

Clicking the "Find text" link redirects the user to the **Google** search system website:



And, in most cases, the first link given by Google will refer to the page that is associated with the text containing the sentence in question:

The screenshot shows a web browser window with the URL `musulman.su/?p=6523`. The main content area displays the title **ИЗГЕ КОРЪӘННЕҢ ТӘҮГЕ БАСМАСЫ 1537 ЕЛДА ВЕНЕЦИЯДӘ ДӨНЬЯ КҮРҮЕ ТУРЫНДА ИШЕТКӘН ИДЕГЕЗМЕ?** and a search bar with the text `Изге Китапның чираттаг`. The search results at the bottom show a single entry: `Изге Китапның чираттаг` with a score of `1-е из 1 совпадения`. The browser interface includes tabs, a search bar, and a sidebar with sections for **Эзләү**, **Радио**, and **ВИДЕО**.

Please note, however, that the sentence will not be highlighted in color. The user can find the sentence by using the search function of the web browser (usually by pressing simultaneously **Ctrl** and **F**), or just looking through the text.

Below the list of example sentences, there are links to other pages, containing more sentences. By clicking "**Next**" the user will be able to see the following 50 examples, and so on, until all sentences containing the targeted word have been displayed.

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left side, there is a navigation menu with links for "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there is a "Contact address" section with the email `tatcorpus@gmail.com` and a "Developers" section listing "Saykhunov M.R." and "Ibragimov T.I.". The main content area displays a list of search results for the word "Китапның". Each result is preceded by a play button icon and includes the source text and a "Find text" link. The results are:

- 1. **Китапның** авторы — алдынгы фикерле гарәп (Мисыр) язучысы (чыгышы белән көрд) Касыйм Әмин (1865—1909).
(source: "Габдулла Тукайга багышланган Интернет-портал" (web-caŭm)) [Find text](#)
- 2. **Һәм** шушы **китапның** дөнья күрүендә Сезнең ярдәм итүегезне сорыйбыз.
(source: "Туганайлар" газетасы (web-caŭm)) [Find text](#)
- 3. **Китапның** беренче өлеше ел фасыллары белән бәйлә горөф-гадәтләрне үз эченә алса, икенче өлешендә гаилә традицияләре каләмгә алынган.
(source: "Белем.ру" порталы (web-caŭm)) [Find text](#)
- 4. **Китапның** беренче кисәгендә 30-35 документ файдаланылган.
(source: РИЗАЭДДИН ФӘХРЕДДИН: МИРАСЫ ҺӘМ ХӘЗЕРГЕ ЗАМАН: Мәкаләләр җыентыгы. Казан, 16 ноябрь 1999 ел. - Казань, 2003.)
- 5. **Китапның** кулъязма вариантларында һәм беренче (1847) басмасында Т. Ялчыголның үзе иҗат иткән байтак шигъри юллар бар.
(source: "Татар әдипләре" порталы (web-caŭm)) [Find text](#)
- 6. Өч **китапның** да үзәген кеше гомере, күңелдәге кичерешләр һәм дөнья турында уйланулар тәшкил иткәнлектән алар бер-берсен тулыландыралар.
(source: "Татар әдипләре" порталы (web-caŭm)) [Find text](#)
- 7. **Китапның** төп өлешен архив документлары, ягъни ревизия материаллары алып тора да.
(source: "Азатлык Радиосы" (web-caŭm)) [Find text](#)

At the bottom of the page, there is a navigation bar with three buttons: "< Prev", "Begin", and "Next >". The "Next >" button is highlighted with a red rectangular box. Below the navigation bar, the text "Laboratory of Experimental Phonetics" is visible.

How to find a lexeme in the Corpus?

Due to the presence of morphological marking of the Corpus, it is possible to search for examples with different word forms of the specified lexeme.

In order to do that, type the lemma of the word (that is, the basic dictionary form of the word, e.g. *тормыш*, *тутыр*, *матур*) in the text field of the Search page and select the checkbox named "Search in the morphologically annotated corpus for lemma...".

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". The search interface includes a language selector (TAT, РУС, ENG), a search type selector, a search input field containing "китап", and a "Find!" button. The search type "Search in the morphologically annotated corpus for lemma..." is selected. The search results are empty. The page footer mentions "Laboratory of Experimental Phonetics".

TAT **РУС** **ENG** [to main page](#)

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны**, **авылларга**, **килмөдегез**)

Search in the morphologically annotated corpus for lemma (for example, **китап**, **авыл**, **кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап***, **авыл***, ***гез**, ***лөргә**)

Laboratory of Experimental Phonetics

After clicking the "Find!" button, you will see a list of sentences where the searched lemma appears in various grammatical forms.

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left side, there is a navigation menu with links for "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there is a "Contact address:" section with the email `tatcorpus@gmail.com` and a "Developers:" section listing "Saykhunov M.R." and "Ibragimov T.I.". The main content area displays search results for the lemma "kuman". The results are listed as follows:

- LEMMA: kuman**
- ▶ Фатирыннан чыгып киткәндә, шагыйрь миңа "Таңнар, кичәләр" дигән **китабын** бүләк итте.
(source: Бару ИСЛАМ. БИК ГАДИ ДӘ, СЕРЛЕ ДӘ СИН, ТОРМЫШ)
- ▶ Вахтаны кабул итеп алганнан соң, өстәл артына утырып йоклап китмәс өчен бер детектив **kuman** укый башлады.
(source: Бару ИСЛАМ. БИК ГАДИ ДӘ, СЕРЛЕ ДӘ СИН, ТОРМЫШ)
- ▶ Болар хакында 1916 елда Уфада чыккан "Башкиры" дигән **kumanта** урыс этнографы яза.
(source: Илдус Хужҗин. Топонимнар)
- ▶ Казан галимнарыннан Д.Рамазанованың "Формирование татарских говоров юго-западной Башкирии" дигән **китабында** да күп "башкорт" авылларының казан татарларынан формалашканы күренә.
(source: Илдус Хужҗин. Топонимнар)
- ▶ "... А.Әсфәндияровның "История Башкирских сел Пермской и Свердловской областей" дигән **китабы** буенча, безнең ата-бабаларыбыз Башкорт җиреннән чыкканнар, чөнки Өртә районы җирләре элек Башкортостанга кергән булган.
(source: Илдус Хужҗин. Топонимнар)
- ▶ Бу авыл халкы (алар таптәрләр дип язылганнар) 1765 елда Сыскы вулысы башкортларынан жир алганнар дип языла "Западные башкиры" дигән **kumanта**.
(source: Илдус Хужҗин. Топонимнар)

Laboratory of Experimental Phonetics

The system of source indication, as well as the full text viewing possibility, is identical with that of the basic word searches described previously.

How to find all lexemes beginning with a particular prefix in the Corpus?

It is possible to search for lexemes having a particular initial part in the Corpus.

To do this, the user should specify the prefix by writing it in the text field of the search page, and then select the checkbox with the name "**Pattern matching search...**". One must note that minimum length of the prefix is 3 characters.

The prefix must be given as in «**авыл***», i.e. ending in the asterisk character.

CORPUS OF WRITTEN TATAR

TAT РУС ENG to main page

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Search the Corpus of Tatar language

Select the search type and enter the desired word:

авыл* Find!

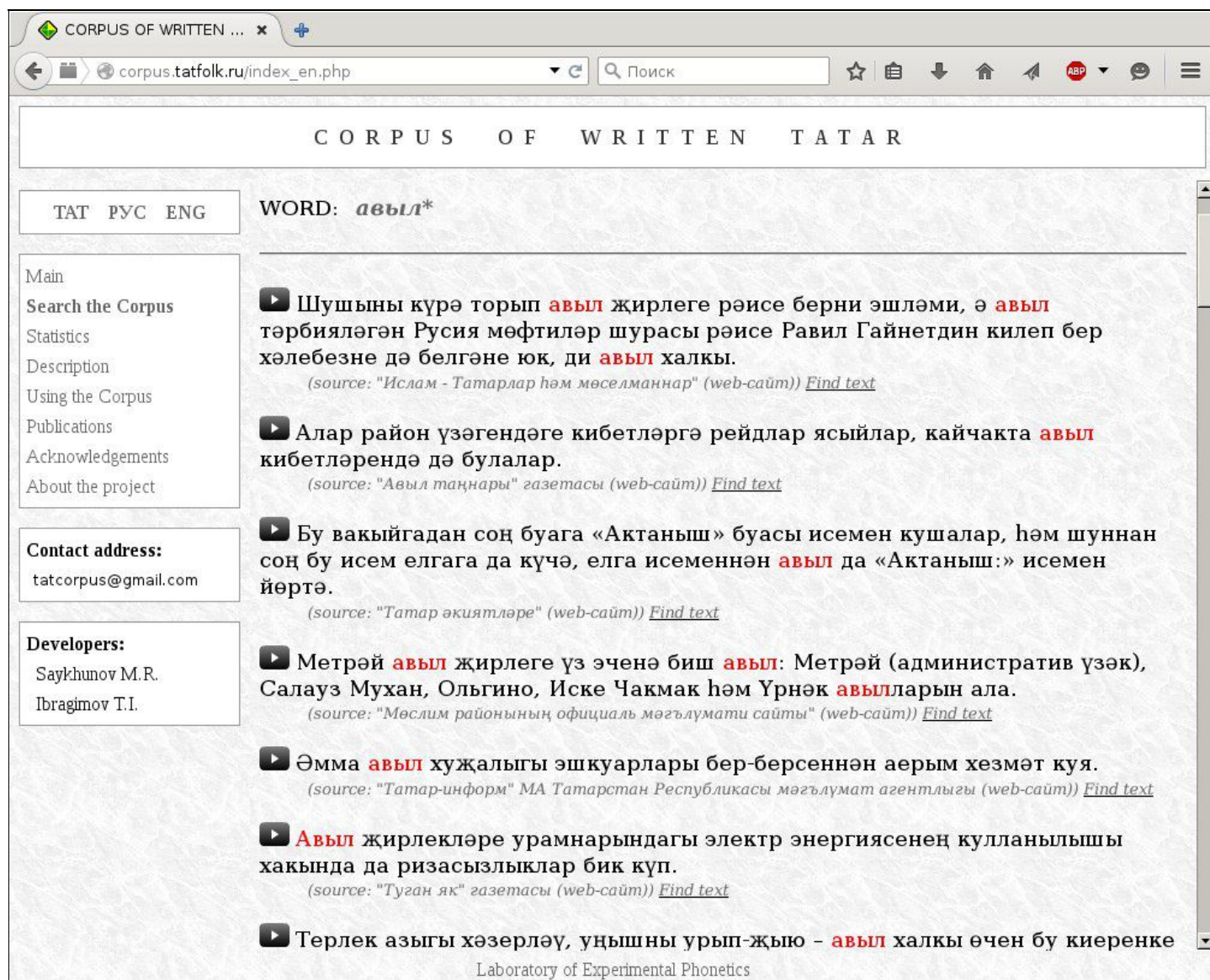
Search for collocations in the contextual (statistic) corpus by wordform (for example, китапны, авылларга, килмәдөгез)

Search in the morphologically annotated corpus for lemma (for example, китап, авыл, кил)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, китап*, авыл*, *гез, *лэргә)

Laboratory of Experimental Phonetics

Here's an example of what you can get as a result of such a query:



The screenshot shows a web browser window with the address bar containing 'corpus.tatfolk.ru/index_en.php'. The page title is 'CORPUS OF WRITTEN TATAR'. Below the title, there are language selection buttons for 'TAT', 'РУС', and 'ENG'. The search results are for the word 'авыл*'. The results are listed in a vertical column, each starting with a play button icon. The first result is: 'Шушыны күрә торып **авыл** жирлеге рәйсе берни эшләми, ә **авыл** тәрбияләгән Русия мөфтиләр шурасы рәйсе Равил Гайнетдин килеп бер хәлебезне дә белгәне юк, ди **авыл** халкы.' with a source note '(source: "Ислам - Татарлар һәм мөселманнар" (web-caŭm)) Find text'. The second result is: 'Алар район үзәгендәге кибетләргә рейдлар ясыйлар, кайчакта **авыл** кибетләрендә дә булалар.' with a source note '(source: "Авыл таңнары" газетасы (web-caŭm)) Find text'. The third result is: 'Бу вакыйгадан соң буага «Актаныш» буасы исемен кушалар, һәм шуннан соң бу исем елгага да күчә, елга исемәннән **авыл** да «Актаныш:» исемән йөртә.' with a source note '(source: "Татар әкиятләре" (web-caŭm)) Find text'. The fourth result is: 'Метрәй **авыл** жирлеге үз эченә биш **авыл**: Метрәй (административ үзәк), Салауз Мухан, Ольгино, Иске Чакмак һәм Үрнәк **авыл**ларын ала.' with a source note '(source: "Мөслим районының официалъ мәгълүмати сайты" (web-caŭm)) Find text'. The fifth result is: 'Әмма **авыл** хужалыгы эшқуарлары бер-берсеннән аерым хезмәт куя.' with a source note '(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caŭm)) Find text'. The sixth result is: '**Авыл** жирлекләре урамнарындагы электр энергиясенә кулланылышы хакында да ризасызлыклар бик күп.' with a source note '(source: "Туган як" газетасы (web-caŭm)) Find text'. The seventh result is: 'Төрлек азыгы хәзерләү, уңышны урып-жыю - **авыл** халкы өчен бу киеренке'. At the bottom of the page, it says 'Laboratory of Experimental Phonetics'.

How to find all lexemes ending in a given postfix in the Corpus?

It is possible to search for lexemes having a particular final part in the Corpus.

To do this, the user should specify the postfix by writing it in the text field of the search page, and then select the checkbox with the name "**Pattern matching search...**". One must note that minimum length of the postfix is 3 characters.

The postfix has to be given as in «*гез», i.e. beginning with the asterisk symbol.

CORPUS OF WRITTEN TATAR

TAT РУС ENG to main page

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Search the Corpus of Tatar language

Select the search type and enter the desired word:

*гез Find!

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)

Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә)**

Laboratory of Experimental Phonetics

Here's an example of what you can get as a result of such a query:

The screenshot shows a web browser window with the address bar containing 'corpus.tatfolk.ru/index_en.php'. The page title is 'CORPUS OF WRITTEN TATAR'. Below the title, there are navigation tabs for 'TAT', 'PVC', and 'ENG'. A search bar contains the word '*gez'. The main content area displays five search results, each starting with a play button icon and followed by a text snippet and a source citation. The results are:

- 1. **Бу ханнар арасындагы көрәш Кече Мөхәммәтнең жиңүе белән төгәлләнә һәм ул Олы Урдага нигез салучы булып санала.**
(source: "Татарча текстлар, программалар" (web-caŭm)) [Find text](#)
- 2. **Узган атнада бәйсез табиблар уздырган экспертиза, журналистны хастаханәдә тоту өчен нигез юк, дип белдергән иде.**
(source: "Азатлык Радиосы" (web-caŭm)) [Find text](#)
- 3. **Чагыштырмача күптән түгел генә нигез салынган мәчетне төзүне тиз тоттылар.**
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caŭm)) [Find text](#)
- 4. **Европа өлгесендәге татар театрына нигез салучы бит син.**
(source: Низами Р.М. Без яшибез... - Казан: Татарстан китап нәшрияты, 2011.)
- 5. **Бүген биредә "Казан заманча төргәкләү заводы"ның нигез ташлары салынды.**
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caŭm)) [Find text](#)

Below the results, there is a sixth entry:

▶ **Безгә крестьян ихтилаллары дип аңлатылган ул вакыттагы Степан Разинның бөтенроссия гражданнар сугышын, чынбарлыкта исә Алтын урданың законлы чын варислары булган Олы урда - Сарай татарлары, (казаклары дип тә әйтергә була) белән Мәскәү арасында Алтын урда империясә тәхетә өчен барган иң канкөешлы, иң аяусыз, иң компромиссыз, иң хәлиткөч соңгы зур алышы булган дип исәпләргә дә тулы нигез бар.**
(source: "Татар,уян!" газетасы (web-caŭm)) [Find text](#)

At the bottom of the page, it says 'Laboratory of Experimental Phonetics'.

How to find a certain word combination in the Corpus?

If you want to find occurrences of a certain word combination (collocation, phrase) in the texts, the contextual-statistical orientation of the Corpus will offer you ways to do the task.

For instance, if you need sentences containing the combination «*китапның эчтәлеген*», you have three ways of finding them:

- 1) **Through the right neighbours.** First, type the word in the left position, i.e. «*китапның*», in the text field of the search page. Then select the checkbox named «**Search for collocations in the contextual (statistic) corpus by wordform...**»

CORPUS OF WRITTEN TATAR

TAT РУС ENG to main page

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Search the Corpus of Tatar language

Select the search type and enter the desired word:

китапның Find!

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)

Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)

Laboratory of Experimental Phonetics

In the window now opened, find the word «эчтәлеген» among the "Right neighbours":

CORPUS OF WRITTEN TATAR

TAT РУС ENG

WORD: *китапның*

NUMBER OF OCCURRENCES: 3120

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

RIGHT NEIGHBOURS: авторы (146; 1625.81), _NUMBER_ (144; 134.06), беренче (119; 579.32), икенче (86; 475.46), исеме (73; 567.74), төп (56; 282.71), исем (55; 453.97), бер (52; 84.03), соңгы (50; 232.22), эчтәлеге (44; 476.6), ахырында (42; 342.94), « (39; 39.47), һәр (39; 152.04), кереш (37; 400.51), “ (30; 3.95), иң (29; 59.37), авторлары (28; 299.1), тиражы (26; 323.12), **эчтәлеген (26; 285.34)**, мөхәррире (26; 226.15), исемен (22; 139.32), эчтәлегенә (21; 262.48), төзүчесе (20; 259.23), баш (17; 47.7), кадрен (16; 134.95), титул (16; 222.34), фәнни (15; 71.71), язылу (14; 117.73), әһәмияте (14; 113.48), дөньяга (13; 62.35), тышлыгы (13; 190.91), тышлыгында (13; 192.15), яңа (13; 10.95), тагын (13; 19.43), электрон (12; 62.8), кыйммәте (12; 122.25), редакторы (12; 109.51), ахырына (12; 77.81), рус (12; 40.54), туыландырылган (11; 145.04), басылып (11; 66.24), үз (11; 4.91), эченә (11; 58.35), башында (11; 48.32), дөнья (11; 28.38), нинди (11; 21.16), берничә (11; 26.21), зурлыгы (11; 110.07), дәвам (10; 89.46), азагында (10; 69.35), икә (10; 7.43), эчәндә (9; 23.93), аерым (9; 21.22), беренчесе (9; 61.76), бәясә (9; 45.17), күп (9; 4.68), тарихы (9; 38.27), үзен (9; 30.2), тышлыгын (9; 143.19), исеменә (9; 52.86), өченчә (9; 31.36), дәрәжәсен (8; 48.51), рәссамы (8; 66.54), үзенә (8; 23.84), чыгуы (8; 57.27), тышлыгына (7; 98.96), идея (7; 47.72), буеннан-буена (7; 78.26), илаһи (7; 49.64), тулы (7; 19.42), тышына (7; 80.96), максаты (7; 28.99), кулъязмасы (7; 74.51), тәржемәсен (7; 82.86), концепциясе (7; 61.05), дүртөнчә (7; 35.75), ролен (6; 34.36), алгы (6; 40.17), берсен (6; 33.5), тәүге (6; 28.69), битләрен (6; 57.76), бәясен (6; 41.32), нигезенә (6; 48.04), урыны (6; 21.71), жаваплы (6; 23.14), нигезен (6; 43.45), авырлыгы (6; 44.28), барлык (6; 7.04), калган (6; 9.22), киңәйтелгән (5; 36.03), битләре (5; 44.65), тәржемәсе (5; 42.59), архив (5; 34.78), озаclamый (5; 31.25), тышында (5; 56.4), баштагы (5; 41.35), кырына (5; 45.43), россиягә (5; 34.23), эчтәлегендә (5; 61.1), рецензенты (5; 86) [Show all](#)

LEFT NEIGHBOURS: %^% (903; 1453.07), бу (305; 1270.4), , (224; 4.69)

Laboratory of Experimental Phonetics

Click this word («**эчтәлеген**»), and see how it turns into highlighted red:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

WORD: **китанның**

NUMBER OF OCCURRENCES: 3120

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

RIGHT NEIGHBOURS: авторы (146; 1625.81), _NUMBER_ (144; 134.06), беренче (119; 579.32), икенче (86; 475.46), исеме (73; 567.74), төп (56; 282.71), исем (55; 453.97), бер (52; 84.03), соңгы (50; 232.22), эчтәлегә (44; 476.6), ахырында (42; 342.94), « (39; 39.47), һәр (39; 152.04), кереш (37; 400.51), “ (30; 3.95), иң (29; 59.37), авторлары (28; 299.1), тиражы (26; 323.12), **эчтәлеген** (26; 285.34), мөхәррире (26; 226.15), исемән (22; 139.32), эчтәлегенә (21; 262.48), төзүчесе (20; 259.23), баш (17; 47.7), кадерен (16; 134.95), титул (16; 222.34), фәнни (15; 71.71), язылу (14; 117.73), әһәмияте (14; 113.48), дөньяга (13; 62.35), тышлыгы (13; 190.91), тышлыгында (13; 192.15), яңа (13; 10.95), тагын (13; 19.43), электрон (12; 62.8), кыйммәте (12; 122.25), редакторы (12; 109.51), ахырына (12; 77.81), рус (12; 40.54), туыландырылган (11; 145.04), басылып (11; 66.24), үз (11; 4.91), эченә (11; 58.35), башында (11; 48.32), дөнья (11; 28.38), нинди (11; 21.16), берничә (11; 26.21), зурлыгы (11; 110.07), дәвам (10; 89.46), азагында (10; 69.35), ике (10; 7.43), эчәндә (9; 23.93), аерым (9; 21.22), беренчесе (9; 61.76), бәясә (9; 45.17), күп (9; 4.68), тарихы (9; 38.27), үзен (9; 30.2), тышлыгын (9; 143.19), исемәнә (9; 52.86), өченчә (9; 31.36), дәрәжәсен (8; 48.51), рәсәмә (8; 66.54), үзенә (8; 23.84), чыгуы (8; 57.27), тышлыгына (7; 98.96), идея (7; 47.72), буеннан-буена (7; 78.26), илаһи (7; 49.64), тулы (7; 19.42), тышына (7; 80.96), максаты (7; 28.99), кулъязмасы (7; 74.51), тәржәмәсен (7; 82.86), концепциясе (7; 61.05), дүртөнчә (7; 35.75), ролен (6; 34.36), алгы (6; 40.17), берсен (6; 33.5), тәүгә (6; 28.69), битләрен (6; 57.76), бәясән (6; 41.32), нигезенә (6; 48.04), урыны (6; 21.71), җаваплы (6; 23.14), нигезен (6; 43.45), авырлыгы (6; 44.28), барлык (6; 7.04), калган (6; 9.22), киңәйтәлгән (5; 36.03), битләре (5; 44.65), тәржәмәсе (5; 42.59), архив (5; 34.78), озакламый (5; 31.25), тышында (5; 56.4), баштагы (5; 41.35), кырыена (5; 45.43), россиягә (5; 34.23), эчтәлегендә (5; 61.1), рецензенты (5; 86) [Show all](#)

LEFT NEIGHBOURS: %^% (903; 1453.07), бу (305; 1270.4), , (224; 4.69)

Laboratory of Experimental Phonetics

Below, in the section "Examples", there will be a list of sentences where this word combination is used:

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left, there is a navigation menu with options: TAT, РУС, ENG; Main; Search the Corpus; Statistics; Description; Using the Corpus; Publications; Acknowledgements; About the project; Contact address: `tatcorpus@gmail.com`; and Developers: Saykhunov M.R., Ibragimov T.I.

The main content area is titled "EXAMPLES:" and contains six entries, each starting with a play button icon and followed by a sentence in Tatar. The phrase "kitapnyñ eçtälegen" is highlighted in red in each sentence. Below each sentence is the source and a "Find text" link.

- ▶ Аджтимушкай турындагы бу язмамның яртысы диярлек Керчътә сатып алынган **kitapnyñ eçtälegen** сөйләп чыгу булуга карамастан, «сөрген» халкына бик сокландыргыч тәэсир ясады ул.
(source: "Идел" журналы (web-caim)) [Find text](#)
- ▶ Элекке хәрбиң адвокаты исә **kitapnyñ eçtälegen** рәсмиләр белән килештерергә кирәклеген кире кага.
(source: "Азатлык Радиосы" (web-caim)) [Find text](#)
- ▶ **Kitapnyñ eçtälegen** бик матур, сәнгатьле сөйли белгән ул.
(source: "Яшел Үзән" газетасы (web-caim)) [Find text](#)
- ▶ Ә Шөгәр төбәгендә гомер иткән Наилә Сиражетдин кызы Бикчуринаның язып калдырган истәлекләре, нәсел шәжәрәсе **kitapnyñ eçtälegen** тарихи әһәмияткә ия мәгълүмат, документлар белән баетты.
(source: "Самара региональ «Дуслык» һәм Самара шәһәр «Ақ бәхет» ижади-иҗтимагый оешмаларының уртак сайты" (web-caim)) [Find text](#)
- ▶ Китап исеме **kitapnyñ eçtälegen** сөйләп бирергә тиеш түгел.
(source: Рабит Батулла. Кәбир Бәкернең тууы)
- ▶ Шулай дип башлап китте дә ул кичә укыган **kitapnyñ eçtälegen** сөйләргә тотынды.
(source: Разил Вәлиев. Кунак булып килде, хужа булып китте...)
- ▶ Әйе, без синең хәтер күзәнәкләренә йөз меңнәрчә **kitapnyñ eçtälegen** язып куйдык.
(source: Адлер Тимергалин. КОХАУ РОНГО-РОНГО)

Laboratory of Experimental Phonetics

2) **Through the left neighbours.** If you use this alternative, write the right-side (latter) word in the text field on the search page, in the given example: «**эчтәлеген**».

CORPUS OF WRITTEN TATAR

TAT РУС ENG [to main page](#)

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

Search the Corpus of Tatar language

Select the search type and enter the desired word:

эчтәлеген Find!

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)

Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)

Laboratory of Experimental Phonetics

Among the "Left neighbours" you will find the word «**китапның**»:

The screenshot shows a web browser window with the address bar containing 'Введите поисковый запрос или адрес' and a search button labeled 'Поиск'. The page title is 'CORPUS OF WRITTEN TATAR'. On the left side, there is a navigation menu with options: 'Main', 'Search the Corpus', 'Statistics', 'Description', 'Using the Corpus', 'Publications', 'Acknowledgements', and 'About the project'. Below the menu, there are sections for 'Contact address: tatcorpus@gmail.com' and 'Developers: Saykhunov M.R., Ibragimov T.I.'. The main content area displays 'RIGHT NEIGHBOURS:' followed by a list of words and their frequencies. The word 'китапның' is highlighted in red in the 'LEFT NEIGHBOURS:' section. At the bottom of the page, it says 'Laboratory of Experimental Phonetics'.

TAT РУС ENG

RIGHT NEIGHBOURS: , (146; 5.21), төшкил (71; 568.22), һәм (65; 60.1), сөйләп (62; 609.69), дә (53; 78.42), яңарту (42; 473.06), сөйләү (38; 452.26), тагын (35; 143.58), ачып (33; 255.32), аңлау (21; 215.15), билгели (19; 171.77), аңлап (17; 129.27), ачу (16; 94.23), тулысынча (16; 101.47), яхшырту (14; 116.92), сөйләргә (14; 120.56), кыскача (14; 130.53), үзгәртү (14; 131.99), баету (14; 169.71), ача (13; 104.56), искә (12; 52.02), сөйли (12; 65.12), баета (11; 129.14), үз (10; 9.06), язып (10; 52.7), ачыклау (9; 65.94), камилләштерү (9; 73.21), дәрәс (9; 29.53), табу (9; 60.23), аңлату (9; 74.93), аңлатып (9; 67.58), үзгәртәргә (8; 67.1), ничек (8; 17.93), аңларга (8; 56.47), үзләштерү (7; 59.95), аңлый (7; 43.1), яхшы (7; 11.76), хасил (7; 58.28), шактый (6; 16.77), төгәл (6; 30.11), бәян (6; 38.33), ачарга (6; 40.23), тирәнәйтәргә (6; 91.36), сөйләгез (6; 66.54), тулырак (6; 46.85), әйтәп (6; 20.64), табабыз (5; 55.02), мәктәптә (5; 19.67), үзәнчәлекләр (5; 23.57), ачуда (5; 45.24), тикшереп (5; 31.86), югалта (5; 40.28), кыска (5; 23.24), аңлауга (5; 59.92), ачар (5; 52.74), баетырга (5; 59.78), язу (5; 27.21), сайлау (5; 18.86), эшләү (5; 19.28), билгеләү (5; 29.75), бермә-бер (5; 41.76), тулыландырып (4; 39.82), бирә (4; 7.02), болай (4; 10.56), ачуга (4; 33.13), язарга (4; 20.58), истәлекләр (4; 30.09), саклаган (4; 30.5), яттан (4; 32.53), үзгәртә (4; 32.33), тулыландыру (4; 38.61), тирән (4; 15.33), саклап (4; 13.36), баетып (4; 42.29), халыкка (4; 13.77), тулы (4; 10.48), табарга (4; 19.72), сөйләтү (4; 60.72), аңлата (4; 19.67), сөйлиләр (4; 27.55), өйрәнә (4; 22.21), белү (4; 20.17), баетуга (4; 46.86), укып (4; 13.67), ачыклай (4; 33.08), төкәдим (4; 5.24), ; (4; 4.68), түбәндәгә (3; 14.01), алдагы (3; 10.51), төлдән (3; 22.33), сөйләде (3; 8.55), үзгәртәп (3; 14.58), беләргә (3; 11.78), яңартуга (3; 27.54), авторның (3; 17.9), микән (3; 11.53), яхшырак (3; 14.17), салырга (3; 14.49), баеткан (3; 34.99), янадан (3; 8.48) [Show all](#)

LEFT NEIGHBOURS: , (179; 26.12), аның (114; 521.98), әсәрнең (91; 1154.86), төп (72; 466.93), һәм (59; 46.83), идея (59; 710.65), кыскача (45; 525.52), текстның (41; 640.25), аларның (38; 173.86), **китапның (26; 285.34)**, әсәрләрнең (25; 312.73), әсәр (25; 215.59), хатның (22; 311.81), бирүнең (21; 235.21), бирү (18; 82.78), тулы (14; 69.61), газетаның (14; 137.22), жырның (13; 146.04), хикәянең (13; 182.25), китапларның (12; 138.96), спектакльнең (11; 116.53)

Laboratory of Experimental Phonetics

After a mouse-click, the word turns red:

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left, there is a navigation menu with options: "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are sections for "Contact address:" (tatcorpus@gmail.com) and "Developers:" (Saykhunov M.R., Ibragimov T.I.).

The main content area displays search results for the word "киташнын". The results are organized into two sections: "RIGHT NEIGHBOURS:" and "LEFT NEIGHBOURS:". In the "RIGHT NEIGHBOURS:" section, the word "киташнын" is highlighted in red. The "LEFT NEIGHBOURS:" section also lists various words and their frequencies.

RIGHT NEIGHBOURS: , (146; 5.21), төшкил (71; 568.22), һәм (65; 60.1), сөйләп (62; 609.69), дә (53; 78.42), яңарту (42; 473.06), сөйләү (38; 452.26), тагын (35; 143.58), ачып (33; 255.32), аңлау (21; 215.15), билгели (19; 171.77), аңлап (17; 129.27), ачу (16; 94.23), тулысынча (16; 101.47), яхшырту (14; 116.92), сөйләргә (14; 120.56), кыскача (14; 130.53), үзгәртү (14; 131.99), баету (14; 169.71), ача (13; 104.56), искә (12; 52.02), сөйли (12; 65.12), баета (11; 129.14), үз (10; 9.06), язып (10; 52.7), ачыклау (9; 65.94), камилләштерү (9; 73.21), дәрәс (9; 29.53), табу (9; 60.23), аңлату (9; 74.93), аңлатып (9; 67.58), үзгәртәргә (8; 67.1), ничек (8; 17.93), аңларга (8; 56.47), үзләштерү (7; 59.95), аңлый (7; 43.1), яхшы (7; 11.76), хасил (7; 58.28), шактый (6; 16.77), төгәл (6; 30.11), бәян (6; 38.33), ачарга (6; 40.23), тирәнәйтәргә (6; 91.36), сөйләгез (6; 66.54), тулырак (6; 46.85), әйтәп (6; 20.64), табабыз (5; 55.02), мәктәптә (5; 19.67), үзәнчәлекле (5; 23.57), ачуда (5; 45.24), тикшерәп (5; 31.86), югалта (5; 40.28), кыска (5; 23.24), аңлауга (5; 59.92), ачар (5; 52.74), баетырга (5; 59.78), язу (5; 27.21), сайлау (5; 18.86), эшләү (5; 19.28), билгеләү (5; 29.75), бермә-бер (5; 41.76), тулыландырып (4; 39.82), бирә (4; 7.02), болай (4; 10.56), ачуга (4; 33.13), язарга (4; 20.58), истәлекләр (4; 30.09), саклаган (4; 30.5), яттан (4; 32.53), үзгәртә (4; 32.33), тулыландыру (4; 38.61), тирән (4; 15.33), саклап (4; 13.36), баетып (4; 42.29), халыкка (4; 13.77), тулы (4; 10.48), табарга (4; 19.72), сөйләтү (4; 60.72), аңлата (4; 19.67), сөйлиләр (4; 27.55), өйрәнә (4; 22.21), белү (4; 20.17), баетуга (4; 46.86), укып (4; 13.67), ачыклай (4; 33.08), төкәдим (4; 5.24), ; (4; 4.68), түбәндәгә (3; 14.01), алдагы (3; 10.51), төлдән (3; 22.33), сөйләде (3; 8.55), үзгәртәп (3; 14.58), беләргә (3; 11.78), яңартуга (3; 27.54), авторның (3; 17.9), микән (3; 11.53), яхшырак (3; 14.17), салырга (3; 14.49), баеткан (3; 34.99), яңадан (3; 8.48) [Show all](#)

LEFT NEIGHBOURS: , (179; 26.12), аның (114; 521.98), әсәрнең (91; 1154.86), төп (72; 466.93), һәм (59; 46.83), идея (59; 710.65), кыскача (45; 525.52), текстның (41; 640.25), аларның (38; 173.86), **киташнын** (26; 285.34), әсәрләренң (25; 312.73), әсәр (25; 215.59), хатның (22; 311.81), бирүнең (21; 235.2), биру (18; 82.78), тулы (14; 69.61), газетаның (14; 137.22), жырның (13; 146.04), хикәянең (13; 182.25), китапларның (12; 138.96), спектакльнең (11; 116.53)

Laboratory of Experimental Phonetics

And, in the section "Examples", you will see the same list of sentences as in the previous, alternative way of searching the word combination in question.

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left side, there is a navigation menu with options: "TAT РУС ENG", "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are sections for "Contact address: tatcorpus@gmail.com" and "Developers: Saykhunov M.R., Ibragimov T.I.". The main content area is titled "EXAMPLES:" and contains seven entries, each starting with a play button icon. Each entry consists of a sentence in Tatar with the phrase "китанның эчтәлеген" highlighted in red, followed by the source and a "Find text" link.

TAT РУС ENG

EXAMPLES:

- ▶ Аджтимушкai турындагы бу язмамның яртысы диярлек Керчътә сатып алынган **китанның эчтәлеген** сөйләп чыгу булуга карамастан, «сөрген» халкына бик сокландыргыч тәэсир ясады ул.
(source: "Идел" журналы (web-caim)) [Find text](#)
- ▶ Элекке хәрбиң адвокаты исә **китанның эчтәлеген** рәсмиләр белән килештерергә кирәклеген кире кага.
(source: "Азатлык Радиосы" (web-caim)) [Find text](#)
- ▶ **Китанның эчтәлеген** бик матур, сәнгатьле сөйли белгән ул.
(source: "Яшел Үзән" газетасы (web-caim)) [Find text](#)
- ▶ Ә Шөгәр төбәгендә гомер иткән Наилә Сиражетдин кызы Бикчуринаның язып калдырган истәлекләре, нәсел шәжәрәсе **китанның эчтәлеген** тарихи әһәмияткә ия мәгълүмат, документлар белән баетты.
(source: "Самара региональ «Дуслык» һәм Самара шәһәр «Ак бәхет» ижади-иҗтимагый оешмаларының уртак сайты" (web-caim)) [Find text](#)
- ▶ Китап исеме **китанның эчтәлеген** сөйләп бирергә тиеш түгел.
(source: Рабит Батулла. Кәбир Бәкернең тууы)
- ▶ Шулай дип башлап китте дә ул кичә укыган **китанның эчтәлеген** сөйләргә тотынды.
(source: Разил Вәлиев. Кунак булып килде, хуҗа булып китте...)
- ▶ Әйе, без сиңең хәтер күзәнәкләреңә йөз меңнәрчә **китанның эчтәлеген** язып куйдык.
(source: Адлер Тимергалин. КОХАУ РОНГО-РОНГО)

Laboratory of Experimental Phonetics

3) **Complex morphological search system.** If you don't need any statistical data and you just want to see examples (sentences) with certain word combinations, you can also use the [Complex morphological search system](#) described in the next chapter.

The Complex morphological search system:

About this system

The morphological marking of the Corpus was carried out in 2014. The meta-language of grammatical labels is based on the system of tags for Turkic languages developed by the international project [Apertium](#). This project aims to develop automatic translating system for a big variety of languages. The main arguments in favor of choosing Apertium's morphological tagger for marking of the Corpus are:

- high quality of morphological annotation;
- Its being Open Source project: all source code and data are publicly available for all for free.

The Complex morphological search system developed by us in 2015-2016 allows one to perform search in the Corpus by different combinations of such parameters as word form, lemma, morphological (grammatical) tags set, beginning of the word, middle part, end of the word, and pointing (indicating) possible distances between the words in question. The maximum length of the search query is five tokens + accordingly four distance specifications between them.

You can open the main query form of the Complex morphological search by clicking on «Search the Corpus» in the main menu and then choosing «**Complex morphological search**» in the select box:

CORPUS OF WRITTEN TATAR

TAT PYC ENG

to main page

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	<input type="text"/>		Distance 1:	<input type="text" value="1-1"/>
Word 2:	<input type="text"/>		Distance 2:	<input type="text" value="1-1"/>
Word 3:	<input type="text"/>		Distance 3:	<input type="text" value="1-1"/>
Word 4:	<input type="text"/>		Distance 4:	<input type="text" value="1-1"/>
Word 5:	<input type="text"/>		<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)
- Complex morphological search (Instruction)**

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

Laboratory of Experimental Phonetics

Let us make a search by the combination «**кунакка бардык**» (“*we went (visited) somewhere as quests*”):

CORPUS OF WRITTEN TATAR

TAT РУС ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	кунакка		Distance 1:	1-1
Word 2:	бардык		Distance 2:	1-1
Word 3:			Distance 3:	1-1
Word 4:			Distance 4:	1-1
Word 5:			<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)
- Complex morphological search (Instruction)

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

Laboratory of Experimental Phonetics

The search parameters are displayed on the top of the main window of search results. Under it, you can find the number occurrences found in the Corpus.

CORPUS OF WRITTEN TATAR

TAT РУС ENG

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications










Acknowledgements

About the project

Useful information

QUERY: {кунакка} 1-1 {бардык} 1-1 {} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 20

-  Икенче көн дә күтәренке күңел белән узды: хатынның туганнарына **кунакка бардык**, яисә үзләре безгә килделәр.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)
-  Менә без дә Чаллыга **кунакка бардык**, шәһәр карап, аның матурлыгына сокланып йөрдек.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)
-  Шактый еллар узгач, аның илле яшьлек юбилеена **кунакка бардык**.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)
-  – Кунаклар килде, **кунакка бардык**, – дип җавап кайтаралар.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)
-  Иртүк мөезин хатынына **кунакка бардык**.
(source: "Идел" журналы (web-caim)) [Find text](#)
-  Без әни белән Нәфисә апаларга **кунакка бардык**.
(source: "Яңарыш" газетасы (web-caim)) [Find text](#)
-  Бервакыт Украинада яшәүче әтиемнең туганнарына **кунакка бардык**.
(source: "Яңарыш" газетасы (web-caim)) [Find text](#)
-  Мәсәлән, узган атнада җәйге лагерыбызда тәрбияләнүче балаларыбыз белән аларга **кунакка бардык**.
(source: "Азатлык Радиосы" (web-caim)) [Find text](#)
-  Без, делегация белән, Яр Чаллы каласына Советлар Союзы Герое Илдар абый Маннанов янына **кунакка бардык**.
(source: "Әдилләр. Әхмәт Дусайлының шәхси порталы" (web-caim)) [Find text](#)

Laboratory of Experimental Phonetics

The search parameters in found sentences are highlighted in red. If you point by the cursor (without clicking) to any of them, a textbox will pop up, indicating the lemma of this word (the basic form of the word in Apertium's interpretation) and a set of morphological tags associated with this word.

C O R P U S O F W R I T T E N T A T A R

TAT РУС ENG

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Useful information

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

QUERY: {кунакка} 1-1 {бардык} 1-1 {} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 20

[▶](#) Икенче көн дә күтәренке күңел белән узды: хатынның туганнарына **кунакка бардык**, яисә үзләре безгә килделәр.
(source: "Безнең гәҗит" газетасы (web-caim)) Find text

[▶](#) Менә без дә Чаллыга **кунакка бардык**, шәһәр карап, аның матурлыгына сокланып йөрдек.
(source: "Безнең гәҗит" газетасы (web-caim)) Find text

[▶](#) Шактый еллар узгач, аның илле яшьлек юбилеена **кунакка бардык**.
(source: "Безнең гәҗит" газетасы (web-caim)) Find text

(кунак) <dat>,<n>,<sg>

[▶](#) – Кунаклар килде, **кунакка бардык**, – дип жавап кайтаралар.
(source: "Безнең гәҗит" газетасы (web-caim)) Find text

[▶](#) Иртүк мөезин хатынына **кунакка бардык**.
(source: "Идел" журналы (web-caim)) Find text

[▶](#) Без әни белән Нәфисә апаларга **кунакка бардык**.
(source: "Яңарыш" газетасы (web-caim)) Find text

[▶](#) Бервакыт Украинада яшәүче әтиемнең туганнарына **кунакка бардык**.
(source: "Яңарыш" газетасы (web-caim)) Find text

[▶](#) Мәсәлән, узган атнада жәйге лагерьыбызда тәрбияләнүче балаларыбыз белән аларга **кунакка бардык**.
(source: "Азатлык Радиосы" (web-caim)) Find text

[▶](#) Без, делегация белән, Яр Чаллы каласына Советлар Союзы Герое Илдар абый Маннанов янына **кунакка бардык**.
(source: "Әдилләр. Әхмәт Дусайлының шәхси порталы" (web-caim)) Find text

Laboratory of Experimental Phonetics

For instance, the text box appearing after moving the cursor to «бардык» (“we went”) of the third sentence:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project
Useful information

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

QUERY: {кунакка} 1-1 {бардык} 1-1 {} 1-1 {} 1-1 {}
NUMBER OF OCCURRENCES: 20

▶ Икенче көн дә күтәренке күңел белән узды: хатынның туганнарына **кунакка бардык**, яисә үзләре безгә килделәр.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)

▶ Менә без дә Чаллыга **кунакка бардык**, шәһәр карап, аның матурлыгына сокланып йөрдек.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)

▶ Шактый еллар узгач, аның илле яшьлек юбилеена **кунакка бардык**.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)

▶ – Кунаклар килде, **кунакка бардык**, – дип җавап кайтаралар.
(source: "Безнең гәҗит" газетасы (web-caim)) [Find text](#)

▶ Иртүк мөезин хатынына **кунакка бардык**.
(source: "Идел" журналы (web-caim)) [Find text](#)

▶ Без әни белән Нәфисә апаларга **кунакка бардык**.
(source: "Яңарыш" газетасы (web-caim)) [Find text](#)

▶ Бервакыт Украинада яшәүче әтиемнең туганнарына **кунакка бардык**.
(source: "Яңарыш" газетасы (web-caim)) [Find text](#)

▶ Мәсәлән, узган атнада җәйге лагерыбызда тәрбияләнүче балаларыбыз белән аларга **кунакка бардык**.
(source: "Азатлык Радиосы" (web-caim)) [Find text](#)

▶ Без, делегация белән, Яр Чаллы каласына Советлар Союзы Герое Илдар абый Маннанов янына **кунакка бардык**.
(source: "Әдилләр. Әхмәт Дусайлының шәхси порталы" (web-caim)) [Find text](#)

Laboratory of Experimental Phonetics

A note for the user: there might also be erroneous tags because the morphological annotation of the Corpus was made automatically.

How to find combinations of defined word forms?

In order to find word combinations in the Corpus, you need to type those words in the right order into the fields «Word 1», «Word 2», «Word 3», «Word 4», «Word 5» one by one. Thus, the maximum possible length of search string is five words.

CORPUS OF WRITTEN TATAR

TAT РУС ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications






Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	<input type="text" value="яңа"/>		Distance 1:	<input type="text" value="1-1"/>
Word 2:	<input type="text" value="ел"/>		Distance 2:	<input type="text" value="1-1"/>
Word 3:	<input type="text" value="бәйрәмендә"/>		Distance 3:	<input type="text" value="1-1"/>
Word 4:	<input type="text"/>		Distance 4:	<input type="text" value="1-1"/>
Word 5:	<input type="text"/>		<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдәгез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)
- Complex morphological search (Instruction)

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

Laboratory of Experimental Phonetics










The results of the search for «яңа ел бәйрәмендә»:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

QUERY: {яңа} 1-1 {ел} 1-1 {бәйрәмендә} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 55

-  Алар **Яңа ел бәйрәмендә** балаларны һәм, гомумән, шундагы һәркемне чын Кыш бабай һәм Кар кызы белән очрашу оештыралар.
(source: "Безнең гәҗит" газетасы (web-сайт)) [Find text](#)
-  Инде китү ягына борылгач та, Наилә апа артымнан килеп: "Кызым, тагын кил. **Яңа ел бәйрәмендә** бик күңелле була", – дип кочып алды.
(source: "Безнең гәҗит" газетасы (web-сайт)) [Find text](#)
-  Бу – **Яңа ел бәйрәмендә** 600 млн литр исерткеч елга булып агачак дигән сүз.
(source: "Безнең гәҗит" газетасы (web-сайт)) [Find text](#)
-  **Яңа ел бәйрәмендә** һәм Сабантуйда.
(source: "Безнең гәҗит" газетасы (web-сайт)) [Find text](#)
-  **Яңа ел бәйрәмендә** уңышлы гына чыгыш ясадык.
(source: "Безнең гәҗит" газетасы (web-сайт)) [Find text](#)
-  **Яңа ел бәйрәмендә** бик күңелле була", – дип кочып алды.
(source: "Безнең гәҗит" газетасы (web-сайт)) [Find text](#)
-  **Яңа ел бәйрәмендә**,
(source: "Туган як" газетасы (web-сайт)) [Find text](#)
-  – Эдуард Анатольевич, халыкны **Яңа ел бәйрәмендә** чыршылар белән тәмин итүгә әзерлек мәсьәләсе ничек тора?
(source: "Зәй офыклары" газетасы (web-сайт)) [Find text](#)
-  Республика **Яңа ел бәйрәмендә** катнашучыларны тәбрикләргә Татарстан Президенты Рөстәм Миңнеханов килде.
(source: Дәүләт хакимияте һәм җирле үзидарә органнарының бердәм "Рәсми Татарстан" порталы (web-сайт)) [Find text](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Developers:

Saykhunov M.R.

Ibragimov T.I.

tatcorpus@gmail.com

Laboratory of Experimental Phonetics

If you need to find words placed not just one after another, but divided by any other words, then you can write numbers into the «Distance» field on the right side of word's text field. For example, «1-1», or just «1», means that the following search word must be placed right after the current one; «2-2» or «2» means that there must be exactly one word between them, and so on. If you type «1-5», the system will look for examples, where the search words in the adjacent fields, have one, two, three, or four words between them. The maximum possible distance (measured in words) between search words is not technically limited, but being, in practice, determined by the length of the sentence.

C O R P U S O F W R I T T E N T A T A R

TAT РУС ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	бик		Distance 1:	2-4
Word 2:	алдылар		Distance 2:	1-1
Word 3:			Distance 3:	1-1
Word 4:			Distance 4:	1-1
Word 5:			<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *дәргә**)
- Complex morphological search (Instruction)

Laboratory of Experimental Phonetics

For example, the results of the search query «{бик} 2-4 {алдылар}» begin as follows:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Useful information

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

QUERY: {бик} 2-4 {алдылар} 1-1 {} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 179

-  Японнар тупны һәм уен кырының үзәген **бик** тиз яулап **алдылар**.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Безнең спортчыларыбыз монда яхшы чыгыш ясый, ләкин алар результатларын яхшыртуны күздә тоталар һәм шуның өчен дә монда **бик** шәп тәҗрибә **алдылар**.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Аккредитация үзәгендә безне **бик** ачык каршы **алдылар**, документларны тикшерделәр, банк картасы белән бейджлар тапшырдылар.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Корылманың персоналы күзәтүчеләр төркемен **бик** мөлаем каршы **алдылар** һәм үз объектларындагы функцияләре турында бик теләп сөйләде.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  “Безне монда **бик** әйбәт каршы **алдылар**, хезмәт күрсәтү бик югары дәрәжәдә оештырылган.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  "Кызларга рәхмәт, **бик** яхшы каршы **алдылар**.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Ут Эстафетасының фирма поезды Курск вокзалына килде, монда Бөтендөнья студентлар Уеннарының төп символын мәскәүлеләр фанфаралар тавышы астында **бик** жылы каршы **алдылар**.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Йөзләрчә төмәнлеләр Эстафета колоннасын бар хәрәкәт маршрутында **бик** жылы каршы **алдылар**.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Атташе булу өчен чит телләргә әйбәт белү генә аз, шуңа да Универсиада өчен аларны **бик** җентекләп сайлап **алдылар**.

Laboratory of Experimental Phonetics

How to find combinations using lemmas?

In case the user wants to perform search based on not a given word form, but on all forms of certain lemma, then he/she should put that lemma in parentheses, for example, «*(китан)*». If the system cannot find a lemma corresponding to the word given in parentheses in its database, the given word will be used as word form in the search.

Let us perform the query «**{теге} {(китан)}**» (“*that (book)*”):

C O R P U S O F W R I T T E N T A T A R

TAT РУС ENG

[to main page](#)

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project
Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	<input type="text" value="теге"/>		Distance 1:	<input type="text" value="1-1"/>
Word 2:	<input type="text" value="(китан)"/>		Distance 2:	<input type="text" value="1-1"/>
Word 3:	<input type="text"/>		Distance 3:	<input type="text" value="1-1"/>
Word 4:	<input type="text"/>		Distance 4:	<input type="text" value="1-1"/>
Word 5:	<input type="text"/>		<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдәгез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)
- Complex morphological search (Instruction)

Laboratory of Experimental Phonetics

The results can be seen below:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

QUERY: {теге} 1-1 {китан} 1-1 {} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 22

- ▶ - Булат, **теге китан** сиңдәме әле? - диде Харис.
(source: Заһит Мурсиев. Ачылмаган тәрәзә)
- ▶ Балачагым күз алдыма килде, "Путивльдән Карпатка кадәр" дигән **теге китан та** искә төште.
(source: "Роберт Миңнуллин (официаль сайт)" (web-caim)) [Find text](#)
- ▶ Әмма **теге китан** күңелдән китми генә.
(source: "Роберт Миңнуллин (официаль сайт)" (web-caim)) [Find text](#)
- ▶ Тагын туп-туры **теге китан** торган бүлеккә ашыгасың.
(source: "Роберт Миңнуллин (официаль сайт)" (web-caim)) [Find text](#)
- ▶ Кави, син минем **теге китан**ны укыдыңмы?
(source: Рабит Батулла. Тузга язылган хәлләр (Тарихи шәхесләрбез, замандашларыбыз турында мәзәкләр, хикәяләр, фаҗигале хәбәрләр))
- ▶ Кави, син минем **теге китан**ны укыдыңмы?
(source: Рабит Батулла. Тузга язылган хәлләр (Тарихи шәхесләрбез, замандашларыбыз турында мәзәкләр, хикәяләр, фаҗигале хәбәрләр))
- ▶ Хәер, теге төрек тарихын, **теге китан**ны биш-алты кат укып чыктым.
(source: Гаяз Исхакый. Тормышмы бу?)
- ▶ - Каенанам хатынга, кияү **теге китабын** биреп торсын әле, мин дә укып чыгар идем, дигән.
(source: "Мәдәни җомга" газетасы (web-caim)) [Find text](#)
- ▶ Ярты сәгать тә үтми, томшыгына кыстырып **теге китан** битен алып та килә.
(source: "Татарча текстлар, программалар" (web-caim)) [Find text](#)
- ▶ **Теге китан**, дәфтәрләр энә шул атларның аяк астында тапталып ятканнар.
(source: "Әдилләр. Әхмәт Дусайлының шәхси порталы" (web-caim)) [Find text](#)

Laboratory of Experimental Phonetics

As another example, we make a search with the specifications «**{{(авыл)}} 1-3 {{(кил)}}**» (“(willage) 1-3 (come)”). Here the search is being performed with two lemmas with possible distance between search words up to two words.

CORPUS OF WRITTEN TATAR

TAT
РУС
ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	(авыл)		Distance 1:	1-3
Word 2:	(кил)		Distance 2:	1-1
Word 3:			Distance 3:	1-1
Word 4:			Distance 4:	1-1
Word 5:			<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *дәргә**)
- Complex morphological search (Instruction)

Laboratory of Experimental Phonetics

The results are as follows:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

QUERY: {авыл} 1-3 {кил} 1-1 {} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 4302

- ▶ – Мәсәдә **авылына** язучылар **килгән**, барабызмы?
(source: Бары ИСЛАМ. БИК ГАДИ ДӘ, СЕРЛЕ ДӘ СИН, ТОРМЫШ)
- ▶ Чөнки якын тирә **авылларга килеп** утырган Гәрәйләр, Меңнәр, Кара бәк нәселләренең шәжәрәләрендәге кешеләр исемнәре бу авылларда әле дә байтак.
(source: Илдус Хужин. Топонимнар)
- ▶ Урта Бәяк турында 17 гасыр документларында мәгълүматлар юк, ахрысы **авыл** соңрак барлыкка **килгән**.
(source: Илдус Хужин. Топонимнар)
- ▶ Бу очракта **авылының** исеме **килеп** чыгышы башка күп кенә авылларныкына туры килә, чөнки Кантуган дигән шәхес тарихта бар.
(source: Илдус Хужин. Топонимнар)
- ▶ Әлбәттә, коңгырат кабиләләреннән тыш бу **авылда** борынгыдан **килгән** төрки ырулар яшәгәндер һәм дә 400-500 еллар элек көчле гәрәйләр һәм Кара бәк кабиләләре анда килеп утырып яңа исем биргәннәр булырга тиеш.
(source: Илдус Хужин. Топонимнар)
- ▶ Әйтергә кирәк, урыслар **авылга** соңрак **килеп** урнашкан булырга тиешләр, чөнки алар татар авылы белән аның зираты арасына кереп утырганнар.
(source: Илдус Хужин. Топонимнар)
- ▶ Шулай итеп Уралда кыпчакларның һәм Кара бәк нәселенең бик күп **авыллары** дөньяга **килә**.
(source: Илдус Хужин. Топонимнар)
- ▶ Ә документлар күрсәтүенчә, 1647 елда Өфе елгасы буендагы Өфе Шигер һәм Әртә Шигер **авыллары** барлыкка **килгән**.
(source: Илдус Хужин. Топонимнар)

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project
Useful information

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatorpus@gmail.com

Laboratory of Experimental Phonetics

How to find combinations using morphological (grammatical) tags?

To learn all tags supported by the Corpus, you can use the following links:

1. http://corpus.tatar/index.php?openinframe=manual/tags_uniq.pdf
2. <https://sourceforge.net/p/apertium/svn/61954/tree//branches/turkic/lexc.rkt>

In the search form, one can define one tag or several tags. Only the combination type «and» is allowed, not «or». This means that it is not possible to use definitions consisting of incompatible (exceptive) tags like in «<dat><acc>» (*dative case, accusative case*) in a search targeting occurrences of one lemma.

As an example, we perform a search by the combination «{<adj>} 1-2 {<n><dat>} 1-3 {<v><past>}». It means that the first word should be an *adjective*, and the following word, by a distance of *zero to one* word, should be a *noun* (<n>) in the *dative case* (<dat>), and after it, by distance up to *two* words, there should be a *verb* (<v>) in the «-заH / -эәH / -каH / -кәH» past tense form (<past>).

CORPUS OF WRITTEN TATAR

TAT PYC ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	<adj>	⌨	Distance 1:	1-2
Word 2:	<n><dat>	⌨	Distance 2:	1-3
Word 3:	<v><past>	⌨	Distance 3:	1-1
Word 4:		⌨	Distance 4:	1-1
Word 5:		⌨	<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдөгез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)
- Complex morphological search (Instruction)

Laboratory of Experimental Phonetics

The results of the search begin with the following:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

QUERY: {<adj>} 1-2 {<n>,<dat>} 1-3 {<v>,<past>} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 10594

- ▶ Авылларыбызның килеп чыгуы вакыты турында фикер йөрткәндә, шуларны онытмау мәгъкуль: башкорт галимнәренең язмалары буенча, бөтен татар авыллары да 1650 – 1750 елларда һәм **соңрак барлыкка килгәннәр.**
(source: Илдус Хужин. Топонимнар)
- ▶ Урта Бәяк турында 17 гасыр документларында мәгълүматлар юк, ахрысы авыл **соңрак барлыкка килгән.**
(source: Илдус Хужин. Топонимнар)
- ▶ **Соңрак бу урынга** башкалар да **килгән.**
(source: Илдус Хужин. Топонимнар)
- ▶ Бу авыл башкортлары (ягъни Жирбиләүчеләре, вотчинники) **соңрак** Әртә **заводына** Жирләрен **сатканнар** һәм үзләре шул ук Жирләрдә яшиләр дип язылган бер документта.
(source: Илдус Хужин. Топонимнар)
- ▶ Авыл халкының байтагы күмер яндырып **якындагы заводларга** ташып **сатканнар.**
(source: Илдус Хужин. Топонимнар)
- ▶ **Ак шәлләре** **жиргә** шуып **төшкән...**
(source: Разил Вәлиев. Шигърыләр, Жырлар)
- ▶ 21 яшьлек шагыйрь 1907 елда үз арамьздан чыккан маһир рәссам хакында хыяланганда, андый рәссам инде хәзерге Татарстанның Буа районы (элеккеге Тәтеш өязе) Күл Черкене авылында 1897 елның 22 февралендә Мәхжүбә абыстай һәм Идрис хәзрәт Урманчелар гаиләсендә **якты дөньяга** килгән **булган.**
(source: "Татарстан китап нәширияты" (web-caim)) [Find text](#)
- ▶ Соңрак рус тарихчылары, мәгънәне берәз тарайтып, барлык төркиләрне татар дип атаганнар, **төрки** сүзе **урынына** татар сүзен **кулланганнар.**
(source: "Татарстан китап нәширияты" (web-caim)) [Find text](#)

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Useful information

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

Laboratory of Experimental Phonetics

In the following example, you find a search targeting two adjacent words where the first word should have the tag for the *second person singular possessive* (<px2sg>) and the second word should be in the *accusative case* (<acc>).

CORPUS OF WRITTEN TATAR

TAT РУС ENG

to main page

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	<input type="text" value="<px2sg>"/>		Distance 1:	<input type="text" value="1-1"/>
Word 2:	<input type="text" value="<acc>"/>		Distance 2:	<input type="text" value="1-1"/>
Word 3:	<input type="text"/>		Distance 3:	<input type="text" value="1-1"/>
Word 4:	<input type="text"/>		Distance 4:	<input type="text" value="1-1"/>
Word 5:	<input type="text"/>		<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдегез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *дәргә**)
- Complex morphological search (Instruction)

Developers:

Saykhunov M.R.

Ibragimov T.I.

tatcorpus@gmail.com

Laboratory of Experimental Phonetics

The results can be seen below:

C O R P U S O F W R I T T E N T A T A R

TAT PYC ENG

QUERY: {<px2sg>} 1-1 {<acc>} 1-1 {} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 4954

- ▶ Шуңа күрә, күптән инде икенче халыкка әйләнгән үз **халкыңның эзләрен** эзләү файдасыз.
(source: Илдус Хужин. Топонимнар)
- ▶ Менә шул үзсүзлегең, кеше сүзен, **анаң сүзен** тыңламавың харап итә яздымы үзеңне?
(source: Фәрит Яхин. СЕРЛЕ КАМКА)
- ▶ Була шундый чаклар, син якын **дустыңның ялганын** тыңлыйсың.
(source: "Татарстан китап нәширияты" (web-caim)) [Find text](#)
- ▶ Такта өстенә **җим саласың да ятьмәне** бакчага чыгарып куясың.
(source: "Татарстан китап нәширияты" (web-caim)) [Find text](#)
- ▶ Тик, зинһар өчен, шаярышмагыз анда, рәссам **абыйларыгызының кушканын** төгәл үтәгез!
(source: "Татарстан китап нәширияты" (web-caim)) [Find text](#)
- ▶ - Их сез, яшел чебешләр, яшь **чагыгызының кадерең** белеп калыгыз.
(source: Заһит Мурсиев. Мин сезнең хатыныгыз)
- ▶ – И улым, болары гына аның бик вак мәсьәлә, артык **исең китмәсен**, менә үз көнеңне үзең күрә башлагач, дөнья әле уңы-суңы яңаклар...
(source: РАВИЛ ӘМИРХАН. ӘМИРХАННАР (тәфсилле шәһәрә). КАЗАН – 2005)
- ▶ Дүшәмбе көн безнең уртақ кичәдә өзәлгән **аягыңның тарихын** сөйләрсен, Зәңгәр чишмә буйлары мондый хәлләрне иштергә бик сусаганнар... ярыймы? — дип, миннән кат-кат вәгдәләр алгач, шаулашып чыгып киттеләр.
(source: ГАЛИМҖАН ИБРАҺИМОВ. КЫЗЫЛ ЧӘЧӘКЛӘР)
- ▶ Ә инде семинар актив аралашуны, балалар белән аралашу өчен үз **программаны формалаштыруны**, яңа методлар, технологияләр табуны күздә тотканың, бу мөгаен, педагогика белеменең иң зур нәтижәседер.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)

Laboratory of Experimental Phonetics

How to find combinations using the beginning, the middle and/or the end of words?

In order to search for all words beginning with certain letters you can use a template like «**ки***». In this case, the system will find such words like «**китап, китапны, киштәгә, кит...**».

You can target the middle part of the word by using a query like «***әме***», which produces **керәмен, әмер, үсәме...**

The template for a search targeting the end of the word looks like «***рны**». As a result, we get sentences with «**дусларны, тарны, барны, кулларны...**».

Let us make a search by the parameters «**{ил*} 1-1 {белән}**», which specify that the first word should begin with «**ил**» and the second word is «**белән**» (*with*).

CORPUS OF WRITTEN TATAR

TAT РУС ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	ил*		Distance 1:	1-1
Word 2:	белән		Distance 2:	1-1
Word 3:			Distance 3:	1-1
Word 4:			Distance 4:	1-1
Word 5:				<input type="button" value="Find!"/>

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдәгез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)
- Complex morphological search (Instruction)

Developers:

Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

Laboratory of Experimental Phonetics

We get the following results:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

QUERY: {ил} 1-1 {белән} 1-1 {} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 2695

Main
Search the Corpus
Tatar Text-To-Speech
User's Guide
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project
Useful information

Developers:
Saykhunov M.R.
Ibragimov T.I.
tatcorpus@gmail.com

- ▶ Элеге масштаблы проектны тормышка ашыру Татарстан башкаласын үстерүгә, аның Россия төбәкләре белән элемтәләрен ныгытуга зур этәргеч булды, халыкта патриотизм һәм **ил белән** хорурлык хисе артты.
(source: "2013 Универсиадасы блогы" (web-сайт)) [Find text](#)
- ▶ Яңа ел театральләштерелгән тамашасында барс балаларының ничек итеп Универсиада Утын эзләр төрле континентлар буенча сәяхәт итүләре, чит **илләр белән** танышулары, төрле мажараларга тарулары, төрле **һөнәрләргә өйрәнүләре һәм ниһаять**, Универсиада Утын кулга төшерүләре тасвирланды.
(source: "2013 Универсиадасы блогы" (web-сайт)) [Find text](#)
- ▶ Универсиадада Россия медальләре өчен Кытай, АКШ кебек **илләр белән** көч сынашчак.
(source: "2013 Универсиадасы блогы" (web-сайт)) [Find text](#)
- ▶ Матди стимуллардан тыш, волонтерлар өчен бик киң күңел ачу программасы да әзерләнгән: Универсиада **Илчеләре белән** иртәнге аш һәм очрашулар, һәр кичтә МЕГАда кичәләр һәм Уеннар тәмамланганнан соң бик якты зур йомгаклау бәйрәме.
(source: "2013 Универсиадасы блогы" (web-сайт)) [Find text](#)
- ▶ Бу максат, әлбәттә, башка **илләр белән** очрашуда да кабатлана.
(source: "2013 Универсиадасы блогы" (web-сайт)) [Find text](#)
- ▶ - Универсиада **илчеләре белән** килешүләре ничек бара?
(source: "2013 Универсиадасы блогы" (web-сайт)) [Find text](#)
- ▶ 2013 елгы Универсиаданың кайда үтәчәге турындагы карарны берничә ел элек бик дулкынланып, тын да алмый **илбез белән** көткән идек.
(source: "2013 Универсиадасы блогы" (web-сайт)) [Find text](#)
- ▶ "Казан 2013" волонтерлары өчен Универсиада **Илчеләре** белән күзгә-күз сөйләшү эш сменасына килү кебек үк гади нәрсә: нәкъ менә объектлардагы буш вакытларында егетләргә һәм кызларны теге яки бу спорт

Laboratory of Experimental Phonetics

To make a search by the beginning and end of the word, you should place * mark between them, e.g. «**ал*лар**» produces words like «**алалар, алдылар, алмалар, алмагачлар...**».

For a search by the beginning, middle part and end of the word, you can shape your query string like «**к*әме*н**», resulting findings like **керәмен, каләмен, күләменнән, кияүдәмен...**

While asterisk sign «*» matches zero or more characters, the question mark «?» represents any single character. For example, pattern «**т?з***» will find words like «**тиз, тозны, түзде, тазарды...**», but not «**тигез, тугызны, тәрәзә...**».

All the listed search parameters (word form, lemma, grammatical tags, the beginning, the middle and end of the word) can be combined in different ways. For example, the query «{<prn>} 1-1 {(кеше)} 1-3 {ал*}» targets word combinations where the first word is a *pronoun* (<prn>), the following word is an occurrence of the *lemma* «кеше» (“*man, people*”), and the third word, separated by a distance up to two tokens from the preceding word, is a word *beginning with* «ал».

CORPUS OF WRITTEN TATAR

TAT РУС ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	<input type="text" value="<prn>"/>		Distance 1:	<input type="text" value="1-1"/>
Word 2:	<input type="text" value="(кеше)"/>		Distance 2:	<input type="text" value="1-3"/>
Word 3:	<input type="text" value="ал*"/>		Distance 3:	<input type="text" value="1-1"/>
Word 4:	<input type="text"/>		Distance 4:	<input type="text" value="1-1"/>
Word 5:	<input type="text"/>		<input type="button" value="Find!"/>	

- Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны, авылларга, килмәдәгез**)
- Search in the morphologically annotated corpus for lemma (for example, **китап, авыл, кил**)
- Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап*, авыл*, *гез, *ләргә**)
- Complex morphological search (Instruction)

Laboratory of Experimental Phonetics









The query produces the following results:

CORPUS OF WRITTEN TATAR

TAT РУС ENG

QUERY: {<prn>} 1-1 {кеше} 1-3 {ал} 1-1 {} 1-1 {}

NUMBER OF OCCURRENCES: 514

-  Мин хажи булгач инде, үзең беләсең **ул кешеләрнең** догаларын **алырға** кирәклегә тугрысында.
(source: РАВИЛ ӘМИРХАН. ӘМИРХАННАР (тәфсилле шәһәрә). КАЗАН – 2005)
-  Хәзер инде **мин кешеләрне**, **аларның** үз-үзләрен тотышын өйрәндем.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Үз күзләрең белән күргәннәреңне **шулкадәр кешеләргә** тапшыра **алуыңны** тою - бик тә рәхәт хис икән ул.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  Бүген бөтен Россия буенча факелчылар тарафыннан күтәрәп барылган һәм 30 зур шәһәрне узган Уеннарның төп символы Универсиада авылында урнашкан Универсиада музееда тора һәм **аны кешеләр** якыннан күрә **алалар**.
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  **Мин кешеләргә** ярдәм итә **алам** икән, нигә моннан файдаланмаска?
(source: "2013 Универсиадасы блогы" (web-caim)) [Find text](#)
-  **Андый кеше Аллаһы Тәгалә** тарафыннан бирелә, лидер мохит эчендә тәрбияләнә ала.
(source: "Үзөбез" яңа буын хәрәкәте (web-caim)) [Find text](#)
-  – **Нинди кеше** булсын **алар**?
(source: Жәүдәт Юныс. Сугыштан соң)
-  Дөрөс, ул **болай да кешеләр** белән артык **алыш**-би-реш итми, аралашмый, аңа йомышы төшеп, йә тәрәзә рамнары, йә балта сабы, йә көянтә-мазар ясатырга килүчеләр булса да, үзе беркемгә йөрми, айга бер-ике кибет тирәсен урап кайта да, минем сезгә катнашым юк, дигәндәй, тагын атавына кереп бикләнә.
(source: Разил Вәлиев. Мирас)

Laboratory of Experimental Phonetics

How to find combinations using several parameters for every word?

A combination of parameters (word form, lemma, grammatical tags, the beginning, the middle and end of the word) can be used to define each word in the search.

For example, consider the following situation. We want to find occurrences of the word form «**алма**» (*verb*, meaning *do not take*), but there is the possibility that sentences with the word «**алма**» (*noun*, meaning *apple*) will also be included in the search results. In order to make the system find only cases where the word «**алма**» is a verb, it is necessary to put the tag «**<v>**», defining the word as a *verb*, immediately after (or before) the word form: «**алма<v>**».

CORPUS OF WRITTEN TATAR

TAT РУС ENGto main page

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1: <input style="border: 2px solid red;" type="text" value="алма<v>"/>			Distance 1: <input type="text" value="1-1"/>
Word 2: <input type="text"/>			Distance 2: <input type="text" value="1-1"/>
Word 3: <input type="text"/>			Distance 3: <input type="text" value="1-1"/>
Word 4: <input type="text"/>			Distance 4: <input type="text" value="1-1"/>
Word 5: <input type="text"/>			<input type="button" value="Find!"/>

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны**, **авылларга**, **килмәдегез**)

Search in the morphologically annotated corpus for lemma (for example, **китап**, **авыл**, **кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап***, **авыл***, ***гез**, ***лөргө**)

Complex morphological search (Instruction, List of tags)

Laboratory of Experimental Phonetics

In this case, there is another way to solve the problem. You can place the right lemma, i.e. «(ал)» (*to take*), immediately after (or before) the word form «алма», resulting in the search definition «алма(ал)». So the system will search for those cases of «алма» where the lemma of the word form is «ал» (*to take*), thus omitting from the results such cases where the lemma is «алма» (*apple*).

CORPUS OF WRITTEN TATAR

TAT РУС ENG

[to main page](#)

Main
[Search the Corpus](#)
[Tatar Text-To-Speech](#)
[User's Guide](#)
[Statistics](#)
[Description](#)
[Using the Corpus](#)
[Publications](#)
[Acknowledgements](#)
[About the project](#)

[Useful information](#)

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1: алма(ал)

Word 2:

Word 3:

Word 4:

Word 5:

Distance 1:

Distance 2:

Distance 3:

Distance 4:

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны**, **авылларга**, **килмәдегез**)

Search in the morphologically annotated corpus for lemma (for example, **китап**, **авыл**, **кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап***, **авыл***, ***гез**, ***ләргә**)

Complex morphological search (Instruction, List of tags)

Laboratory of Experimental Phonetics

Entering search parameters in the graphical mode

A special tool facilitates entering of the search parameters. You can access the tool by clicking the buttons to the right of the text fields, as shown below.

C O R P U S O F W R I T T E N T A T A R

TAT РУС ENG

[to main page](#)

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications











Acknowledgements

About the project

Useful information

Search the Corpus of Tatar language

Select the search type and enter the desired word:

Word 1:	<input type="text"/>	 	Distance 1:	<input type="text" value="1-1"/>
Word 2:	<input type="text"/>	 	Distance 2:	<input type="text" value="1-1"/>
Word 3:	<input type="text"/>	 	Distance 3:	<input type="text" value="1-1"/>
Word 4:	<input type="text"/>	 	Distance 4:	<input type="text" value="1-1"/>
Word 5:	<input type="text"/>	 		<input type="button" value="Find!"/>

Search for collocations in the contextual (statistic) corpus by wordform (for example, **китапны**, **авылларга**, **килмәдегез**)

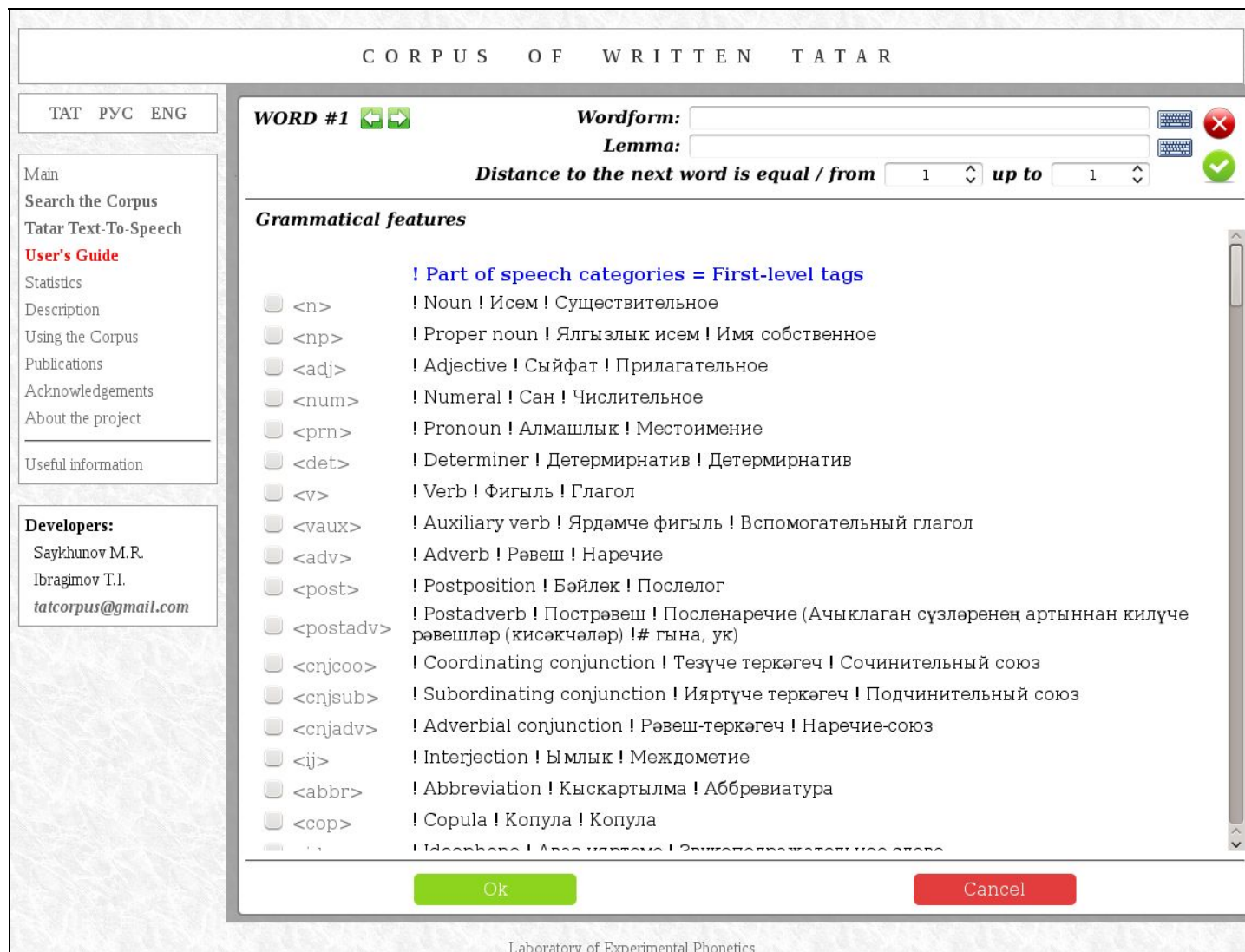
Search in the morphologically annotated corpus for lemma (for example, **китап**, **авыл**, **кил**)

Pattern matching search (the beginning or the end of the word with length 3 and more characters should be specified, for example, **китап***, **авыл***, ***гез**, ***ләргә**)

Complex morphological search (Instruction, List of tags)

Laboratory of Experimental Phonetics

A window of the following shape will pop up.



In the upper left corner, you can see the number of the current word. On the right side of the word, there are two buttons that allow you to quickly jump to editing the previous or the next word. The information entered into the current page will be automatically saved, when you switch between pages!

C O R P U S O F W R I T T E N T A T A R

TAT PYC ENG

WORD #1

Wordform:
Lemma:

Distance to the next word is equal / from *up to*

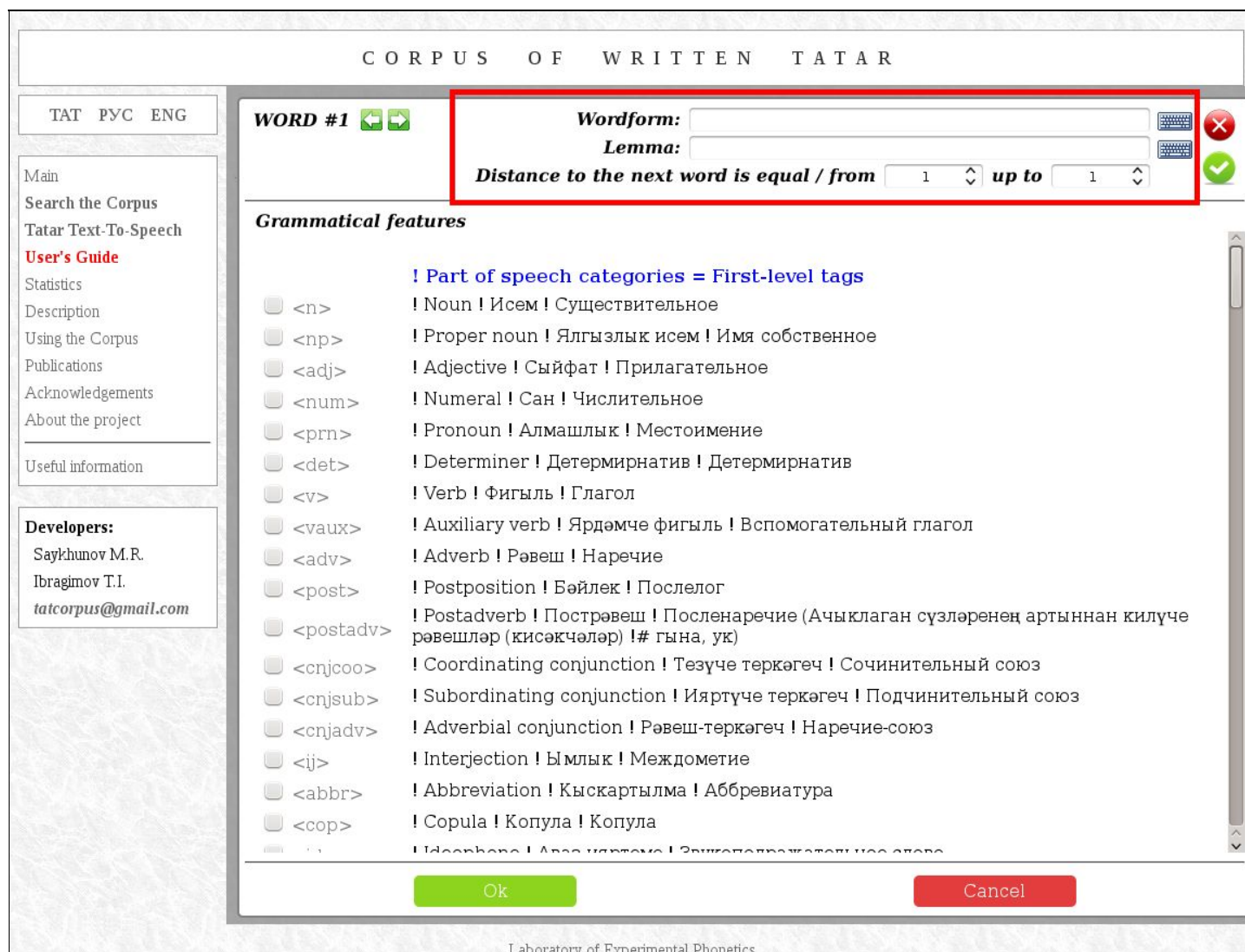
Grammatical features

! Part of speech categories = First-level tags

- <n> ! Noun ! Исем ! Существительное
- <np> ! Proper noun ! Ялгызлык исем ! Имя собственное
- <adj> ! Adjective ! Сыйфат ! Прилагательное
- <num> ! Numeral ! Сан ! Числительное
- <prn> ! Pronoun ! Алмашлык ! Местоимение
- <det> ! Determiner ! Детерминатив ! Детерминатив
- <v> ! Verb ! Фигыль ! Глагол
- <vaux> ! Auxiliary verb ! Ярдәмче фигыль ! Вспомогательный глагол
- <adv> ! Adverb ! Рәвеш ! Наречие
- <post> ! Postposition ! Бәйлек ! Послелог
- <postadv> ! Postadverb ! Пострәвеш ! Посленаречие (Ачыклаган сүزلәренәң артынан килүче рәвешләр (кисәкчәләр) !# гына, ук)
- <cnjcoo> ! Coordinating conjunction ! Тезүче теркәгеч ! Сочинительный союз
- <cnjsub> ! Subordinating conjunction ! Ияртүче теркәгеч ! Подчинительный союз
- <cnjadv> ! Adverbial conjunction ! Рәвеш-теркәгеч ! Наречие-союз
- <ij> ! Interjection ! Ымлык ! Междометие
- <abbr> ! Abbreviation ! Кыскартылма ! Аббревиатура
- <cop> ! Copula ! Копула ! Копула
- <id> ! Ideophone ! Асс истемә ! Звукосопределение слова

Laboratory of Experimental Phonetics

The red highlighted area on the picture that shows where you can specify the word form, lemma and interval. Please note that here the lemma must be entered without parentheses!



The main area of the window (highlighted in red) is where you can choose grammatical features to be included in the search pattern. The tags are listed by category. You can choose a particular tag by checking the box in the corresponding line.

CORPUS OF WRITTEN TATAR

TAT PYC ENG

WORD #1

Wordform:

Lemma:

Distance to the next word is equal / from

up to

Grammatical features

! Part of speech categories = First-level tags

<input type="checkbox"/> <n>	! Noun ! Исем ! Существительное
<input type="checkbox"/> <np>	! Proper noun ! Ялгызык исем ! Имя собственное
<input type="checkbox"/> <adj>	! Adjective ! Сыйфат ! Прилагательное
<input type="checkbox"/> <num>	! Numeral ! Сан ! Числительное
<input type="checkbox"/> <prn>	! Pronoun ! Алмашлык ! Местоимение
<input type="checkbox"/> <det>	! Determiner ! Детерминатив ! Детерминатив
<input type="checkbox"/> <v>	! Verb ! Фигыль ! Глагол
<input type="checkbox"/> <vaux>	! Auxiliary verb ! Ярдәмче фигыль ! Вспомогательный глагол
<input type="checkbox"/> <adv>	! Adverb ! Рәвеш ! Наречие
<input type="checkbox"/> <post>	! Postposition ! Бәйләк ! Послелог
<input type="checkbox"/> <postadv>	! Postadverb ! Пострәвеш ! Посленаречие (Ачыклаган сүзләренәң артыннан килүче рәвешләр (кисәкчәләр) !# гына, ук)
<input type="checkbox"/> <cnjcoo>	! Coordinating conjunction ! Тезүче теркәгеч ! Сочинительный союз
<input type="checkbox"/> <cnjsub>	! Subordinating conjunction ! Ияртүче теркәгеч ! Подчинительный союз
<input type="checkbox"/> <cnjadv>	! Adverbial conjunction ! Рәвеш-теркәгеч ! Наречие-союз
<input type="checkbox"/> <ij>	! Interjection ! Ымлык ! Междометие
<input type="checkbox"/> <abbr>	! Abbreviation ! Кыскартылма ! Аббревиатура
<input type="checkbox"/> <cop>	! Copula ! Копула ! Копула
<input type="checkbox"/> <idphn>	! Ideophone ! Аңаң идрәме ! Звукосопоставительное слово

Laboratory of Experimental Phonetics

The buttons in the upper right corner and in the bottom of the page allow you to close the window with saving (green) or without saving (red) the changes.

CORPUS OF WRITTEN TATAR

TAT РУС ENG

WORD #1

Wordform:

Lemma:

Distance to the next word is equal / from 1 up to 1

Grammatical features

! Part of speech categories = First-level tags

- <n> ! Noun ! Исем ! Существительное
- <np> ! Proper noun ! Ялгызлык исем ! Имя собственное
- <adj> ! Adjective ! Сыйфат ! Прилагательное
- <num> ! Numeral ! Сан ! Числительное
- <prn> ! Pronoun ! Алмашлык ! Местоимение
- <det> ! Determiner ! Детерминатив ! Детерминатив
- <v> ! Verb ! Фигыль ! Глагол
- <vaux> ! Auxiliary verb ! Ярдәмче фигыль ! Вспомогательный глагол
- <adv> ! Adverb ! Рәвеш ! Наречие
- <post> ! Postposition ! Бәйләк ! Послелог
- <postadv> ! Postadverb ! Пострәвеш ! Посленаречие (Ачыклаган сүзләренәң артыннан килүче рәвешләр (кисәкчәләр) !# гына, ук)
- <cnjcoo> ! Coordinating conjunction ! Тезүче теркәгеч ! Сочинительный союз
- <cnjsub> ! Subordinating conjunction ! Ияртүче теркәгеч ! Подчинительный союз
- <cnjadv> ! Adverbial conjunction ! Рәвеш-теркәгеч ! Наречие-союз
- <ij> ! Interjection ! Ымлык ! Междометие
- <abbr> ! Abbreviation ! Кыскартылма ! Аббревиатура
- <cop> ! Copula ! Копула ! Копула

Ok Cancel

Laboratory of Experimental Phonetics

As an example, consider a search for the word form «алма» (*noun*). As the first step, enter the word «алма» into the word form field, and then, select the appropriate line in the tag list.

CORPUS OF WRITTEN TATAR

TAT РУС ENG

WORD #1

Wordform: алма

Lemma:

Distance to the next word is equal / from 1 up to 1

Grammatical features

! Part of speech categories = First-level tags

- <n> ! Noun ! Исем ! Существительное
- <np> ! Proper noun ! Ялгызлык исем ! Имя собственное
- <adj> ! Adjective ! Сыйфат ! Прилагательное
- <num> ! Numeral ! Сан ! Числительное
- <prn> ! Pronoun ! Алмашлык ! Местоимение
- <det> ! Determiner ! Детерминатив ! Детерминатив
- <v> ! Verb ! Фигыль ! Глагол
- <vaux> ! Auxiliary verb ! Ярдәмче фигыль ! Вспомогательный глагол
- <adv> ! Adverb ! Рәвеш ! Наречие
- <post> ! Postposition ! Бәйләк ! Послелог
- <postadv> ! Postadverb ! Пострәвеш ! Посленаречие (Ачыклаган сүзләренәң артыннан килүче рәвешләр (кисәкчәләр) !# гына, ук)
- <cnjcoo> ! Coordinating conjunction ! Тезүче теркәгеч ! Сочинительный союз
- <cnjsub> ! Subordinating conjunction ! Ияртүче теркәгеч ! Подчинительный союз
- <cnjadv> ! Adverbial conjunction ! Рәвеш-теркәгеч ! Наречие-союз
- <ij> ! Interjection ! Ымлык ! Междометие
- <abbr> ! Abbreviation ! Кыскартылма ! Аббревиатура
- <cop> ! Copula ! Копула ! Копула
- <phn> ! Ideophone ! Арау ияремә ! Звукосопыраштыру сүзләре

Ok Cancel

Laboratory of Experimental Phonetics

Another way of defining a search that targets the same word forms is by specifying «алма» as the word form and «алма» as the lemma. In this case, the system will drop occurrences of the verb «алма» from search results because its lemma is «ал».

CORPUS OF WRITTEN TATAR

TAT РУС ENG

Main

Search the Corpus

Tatar Text-To-Speech

User's Guide

Statistics

Description

Using the Corpus

Publications

Acknowledgements

About the project

Useful information

Developers:

Saykhunov M.R.

Ibragimov T.I.

tatcorpus@gmail.com

WORD #1

Wordform:

Lemma:

Distance to the next word is equal / from *up to*

Grammatical features

! Part of speech categories = First-level tags

- <n> ! Noun ! Исем ! Существительное
- <np> ! Proper noun ! Ялгызлык исем ! Имя собственное
- <adj> ! Adjective ! Сыйфат ! Прилагательное
- <num> ! Numeral ! Сан ! Числительное
- <prn> ! Pronoun ! Алмашлык ! Местоимение
- <det> ! Determiner ! Детерминатив ! Детерминатив
- <v> ! Verb ! Фигыль ! Глагол
- <vaux> ! Auxiliary verb ! Ярдәмче фигыль ! Вспомогательный глагол
- <adv> ! Adverb ! Рәвеш ! Наречие
- <post> ! Postposition ! Бәйләк ! Послелог
- <postadv> ! Postadverb ! Пострәвеш ! Посленаречие (Ачыклаган сүзләренәң артыннан килүче рәвешләр (кисәкчәләр) !# гына, ук)
- <cnjcoo> ! Coordinating conjunction ! Тезүче теркәгеч ! Сочинительный союз
- <cnjsub> ! Subordinating conjunction ! Ияртүче теркәгеч ! Подчинительный союз
- <cnjadv> ! Adverbial conjunction ! Рәвеш-теркәгеч ! Наречие-союз
- <ij> ! Interjection ! Ымлык ! Междометие
- <abbr> ! Abbreviation ! Кыскартылма ! Аббревиатура
- <cop> ! Copula ! Копула ! Копула
- <idphn> ! Ideophone ! Ассонанс ! Звукосопоставуучу сүзләр

Ok
Cancel

Laboratory of Experimental Phonetics

As yet another example, consider the task of finding verbs (<v>) starting with a given combination of letters («ac*»). This request should be set in the graphical mode as follows:

CORPUS OF WRITTEN TATAR

TAT PYC ENG

WORD #1

Wordform:

Lemma:

Word begins with:

Word ends with:

Distance to the next word is equal / from up to

Grammatical features

! Part of speech categories = First-level tags

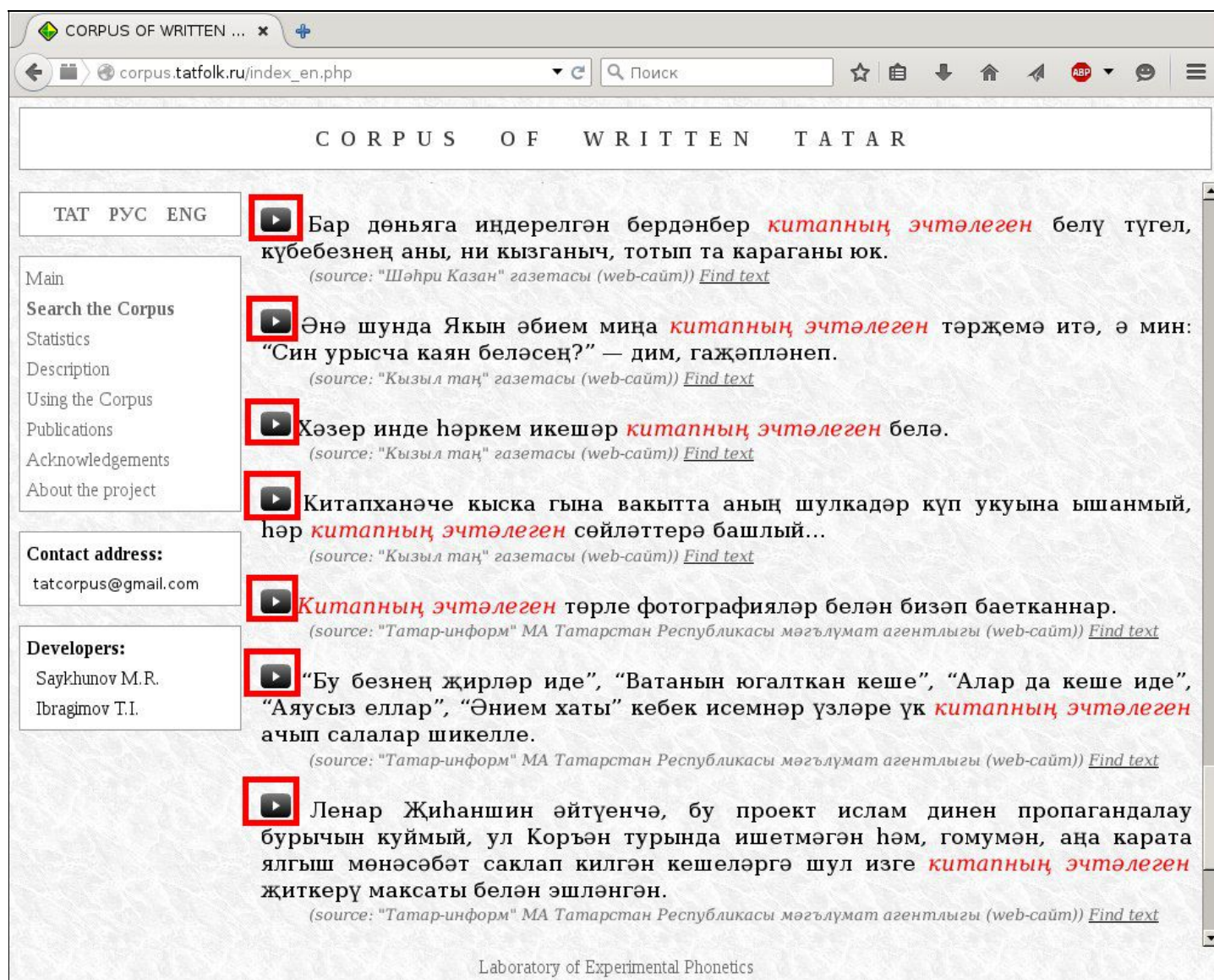
- <n> ! Noun ! Исем ! Существительное
- <np> ! Proper noun ! Ялгызлык исем ! Имя собственное
- <adj> ! Adjective ! Сыйфат ! Прилагательное
- <num> ! Numeral ! Сан ! Числительное
- <prn> ! Pronoun ! Алмашлык ! Местоимение
- <det> ! Determiner ! Детерминатив ! Детерминатив
- <v> ! Verb ! Фигыль ! Глагол**
- <vaux> ! Auxiliary verb ! Ярдәмче фигыль ! Вспомогательный глагол
- <adv> ! Adverb ! Рәвеш ! Наречие
- <post> ! Postposition ! Бәйләк ! Послелог
- <postadv> ! Postadverb ! Пострәвеш ! Посленаречие (Ачыклаган сүзләренәң артыннан килүче рәвешләр (кисәкчәләр) !# гына, ук)
- <cnjcoo> ! Coordinating conjunction ! Тезүче теркәгеч ! Сочинительный союз
- <cnjsub> ! Subordinating conjunction ! Ияртүче теркәгеч ! Подчинительный союз

Ok Cancel

Laboratory of Experimental Phonetics

How to listen to the sentences found in the Corpus?

The Corpus of Written Tatar offers the user a unique opportunity to listen to the sentences found in a search! To do this, just click the special button placed to the left of each sentence.



The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left side, there is a navigation menu with options: "TAT РУС ENG", "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are sections for "Contact address: tatcorpus@gmail.com" and "Developers: Saykhunov M.R., Ibragimov T.I.". The main content area displays a list of search results. Each result consists of a sentence in Tatar, a source attribution, and a "Find text" link. A red square with a play button icon is placed to the left of each sentence, indicating that the user can click it to listen to the audio of that sentence. The sentences are:

- Бар дөнъяга индерелгән бердәнбер **китапның эчтәлеген** белү түгел, күбебезнең аны, ни кызганыч, тотып та караганы юк.
(source: "Шәһри Казан" газетасы (web-caŭm)) [Find text](#)
- Әнә шунда Якын әбием миңа **китапның эчтәлеген** тәржемә итә, ә мин: "Син урысча каян беләсең?" — дим, гажәплөнөп.
(source: "Кызыл таң" газетасы (web-caŭm)) [Find text](#)
- Хәзер инде һәркем икешәр **китапның эчтәлеген** белә.
(source: "Кызыл таң" газетасы (web-caŭm)) [Find text](#)
- Китапханәче кыска гына вакытта аның шулкадәр күп укуына ышанмый, һәр **китапның эчтәлеген** сөйләттерә башлый...
(source: "Кызыл таң" газетасы (web-caŭm)) [Find text](#)
- Китапның эчтәлеген** төрле фотографияләр белән бизәп баеткаллар.
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caŭm)) [Find text](#)
- "Бу безнең жирләр иде", "Ватанын югалткан кеше", "Алар да кеше иде", "Аяусыз еллар", "Әнием хаты" кебек исемнәр үзләре үк **китапның эчтәлеген** ачып салалар шикелле.
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caŭm)) [Find text](#)
- Ленар Жиһаншин әйтүенчә, бу проект ислам динен пропагандалау бурычын куймый, ул Коръән турында ишетмәгән һәм, гомумән, аңа карата ялгыш мөнәсәбәт саклап килгән кешеләргә шул изге **китапның эчтәлеген** житкерү максаты белән эшләнгән.
(source: "Татар-информ" МА Татарстан Республикасы мәгълүмат агентлыгы (web-caŭm)) [Find text](#)

Laboratory of Experimental Phonetics

Additional statistical materials!

The creators of the Corpus upload various additional statistical data as soon as they become available as a result of processing the Corpus. To date, these include:

- A list of the most frequent word forms of the Tatar language.
- A list of the most frequent collocations consisting of 2, 3, 4, 5 and 6 words.
- Frequency lists of letters and their combinations.

The screenshot shows a web browser window with the address bar containing 'CORPUS OF WRITTEN ...'. The page title is 'CORPUS OF WRITTEN TATAR'. Below the title, there are three tabs: 'TAT', 'РУС', and 'ENG'. The main content area is titled 'Statistics' and contains the following information:

Size of the Corpus of Written Tatar: over 116 mln. tokens.
Amount of sentences in the database is more than 10 mln.

Words:
The most frequent 200 wordforms of Tatar language.
The most frequent 200 "2-grams" (wordforms) of Tatar language.
The most frequent 200 "3-grams" (wordforms) of Tatar language.
The most frequent 200 "4-grams" (wordforms) of Tatar language.
The most frequent 200 "5-grams" (wordforms) of Tatar language.
The most frequent 200 "6-grams" (wordforms) of Tatar language.

Letters:
The list of frequency of letters in Tatar language.
The most frequent 200 "2-grams" (letters) of Tatar language.
The most frequent 200 "3-grams" (letters) of Tatar language.
The most frequent 200 "4-grams" (letters) of Tatar language.

At the bottom of the page, it says 'Laboratory of Experimental Phonetics'.

On the left side, there is a navigation menu with the following items: Main, Search the Corpus (highlighted with a red box), Statistics, Description, Using the Corpus, Publications, Acknowledgements, and About the project. Below the menu, there is a 'Contact address:' section with 'tatcorpus@gmail.com' and a 'Developers:' section with 'Saykhunov M.R.' and 'Ibragimov T.I.'.

- Frequency lists of letters and their combinations in the initial and final positions of words.
- Frequency lists of phonemes and their combinations within a word and rhythmic group.

CORPUS OF WRITTEN TATAR

TAT PYC ENG

Main
Search the Corpus
 Statistics
 Description
 Using the Corpus
 Publications
 Acknowledgements
 About the project

Contact address:
 tatcorpus@gmail.com

Developers:
 Saykhunov M.R.
 Ibragimov T.I.

Letters (at the beginnig of a word):

The most frequent 200 "2-grams" (letters, beginnig of a word) of Tatar language.
 The most frequent 200 "3-grams" (letters, beginnig of a word) of Tatar language.
 The most frequent 200 "4-grams" (letters, beginnig of a word) of Tatar language.
 The most frequent 200 "5-grams" (letters, beginnig of a word) of Tatar language.
 The most frequent 200 "6-grams" (letters, beginnig of a word) of Tatar language.

Letters (at the end of a word):

The most frequent 200 "2-grams" (letters, end of a word) of Tatar language.
 The most frequent 200 "3-grams" (letters, end of a word) of Tatar language.
 The most frequent 200 "4-grams" (letters, end of a word) of Tatar language.
 The most frequent 200 "5-grams" (letters, end of a word) of Tatar language.
 The most frequent 200 "6-grams" (letters, end of a word) of Tatar language.

Phonemes (within a rhythmic group):

The list of frequency of phonemes in Tatar language.
 The most frequent 100 "2-grams" (phonemes) of Tatar language.
 The most frequent 100 "3-grams" (phonemes) of Tatar language.

Laboratory of Experimental Phonetics

For example, the list of the 5000 most frequent word forms of Tatar language is as follows:

The screenshot shows a web browser window with the URL 'CORPUS OF WRITTEN TATAR'. The page features a navigation menu on the left with options like 'Main', 'Search the Corpus', 'Statistics', 'Description', 'Using the Corpus', 'Publications', 'Acknowledgements', and 'About the project'. Below the menu, there is a 'Contact address' section with 'tatcorpus@gmail.com' and a 'Developers' section listing 'Saykhunov M.R.' and 'Ibragimov T.I.'. The main content area displays a table of the 5000 most frequent word forms.

Rank	Word Form	Frequency
1	һәм	1630726
2	белән	1249114
3	да	1204431
4	дә	910669
5	бу	746627
6	ул	727682
7	дип	638244
8	өчен	603250
9	бер	508703
10	иде	417602
11	аның	353914
12	буенча	341434
13	ә	340479
14	татар	309106
15	генә	289810
16	бик	286219
17	инде	283772
18	шул	269870
19	бар	265782
20	түгел	264347
21	гына	264131
22	шулай	257466
23	Татарстан	255036
24	үз	251835
25	ук	251452
26	турында	242908
27	алар	242730
28	булган	242209
29	мин	241080
30	булып	239581
31	ел	232660
32	зур	225474
33	кеше	224115
34	алып	218246
35	иң	217754
36	яңа	213919
37	итеп	207815
38	дәүләт	207369
39	соң	204223
40	кирәк	198707
41	әлеге	198508
42	беренче	196135
43	торган	194072
44	күп	192507
45	то	191750

Laboratory of Experimental Phonetics

The first column of the list shows the rank of the word form, the second gives the word form itself, and in the third column one can see the number of occurrences in the Corpus.

Below, a frequency list of combinations of two letters is shown:

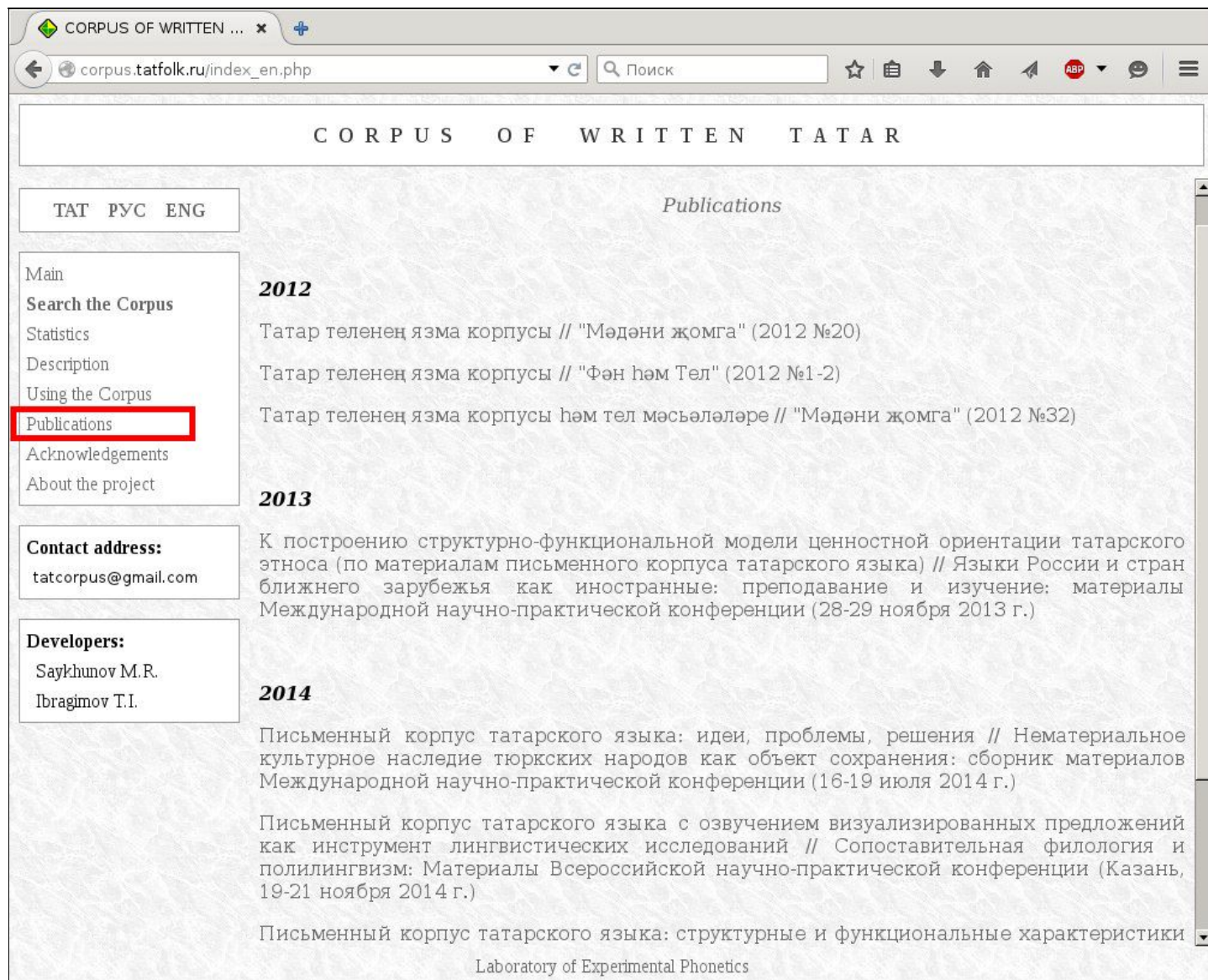
The screenshot shows a web browser window with the title 'CORPUS OF WRITTEN TATAR'. The browser's address bar contains the text 'Введите поисковый запрос или адрес' and a search icon. The website's main header displays 'CORPUS OF WRITTEN TATAR'. On the left side, there is a navigation menu with links for 'Main', 'Search the Corpus', 'Statistics', 'Description', 'Using the Corpus', 'Publications', 'Acknowledgements', and 'About the project'. Below the menu, there is a 'Contact address:' section with the email 'tatcorpus@gmail.com' and a 'Developers:' section listing 'Saykhunov M.R.' and 'Ibragimov T.I.'. The main content area displays a frequency list of two-letter combinations, numbered 1 to 45. The list is as follows:

1	13487484	ар
2	10927856	ла
3	10537847	ан
4	9389724	лә
5	8426490	ен
6	7742249	эр
7	7508939	ал
8	7476037	га
9	7226503	та
10	6947006	ын
11	6721316	да
12	6716396	ел
13	6538509	ка
14	6512923	ер
15	6420902	эн
16	5772159	ны
17	5497572	ре
18	5470991	нд
19	5292593	ры
20	5203472	дә
21	5122055	на
22	5101267	ле
23	5056469	ра
24	4975734	бе
25	4812588	лы
26	4644417	ул
27	4600942	ат
28	4458442	те
29	4039938	ба
30	3887440	не
31	3817973	ма
32	3731737	ты
33	3708165	ге
34	3700021	гә
35	3649670	бу
36	3507257	гы
37	3492769	тә
38	3456093	ин
39	3448221	ак
40	3393397	нә
41	3337527	мә
42	3275056	ки
43	3209149	ит
44	3175581	ыр
45	3146247	ог

At the bottom of the page, the text 'Laboratory of Experimental Phonetics' is visible.

Publications

The "**Publications**" section of the site gives information about all relevant articles published by creators of the Corpus:



The screenshot shows a web browser window with the address bar displaying "corpus.tatfolk.ru/index_en.php". The page title is "CORPUS OF WRITTEN TATAR". The main content area is titled "Publications" and lists articles from 2012, 2013, and 2014. A sidebar on the left contains navigation links, with "Publications" highlighted in a red box. The footer of the page reads "Laboratory of Experimental Phonetics".

TAT PYC ENG

Publications

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

2012

Татар теленең язма корпусы // "Мәдәни жомга" (2012 №20)
Татар теленең язма корпусы // "Фән һәм Тел" (2012 №1-2)
Татар теленең язма корпусы һәм тел мәсьәләләре // "Мәдәни жомга" (2012 №32)

2013

К построению структурно-функциональной модели ценностной ориентации татарского этноса (по материалам письменного корпуса татарского языка) // Языки России и стран ближнего зарубежья как иностранные: преподавание и изучение: материалы Международной научно-практической конференции (28-29 ноября 2013 г.)

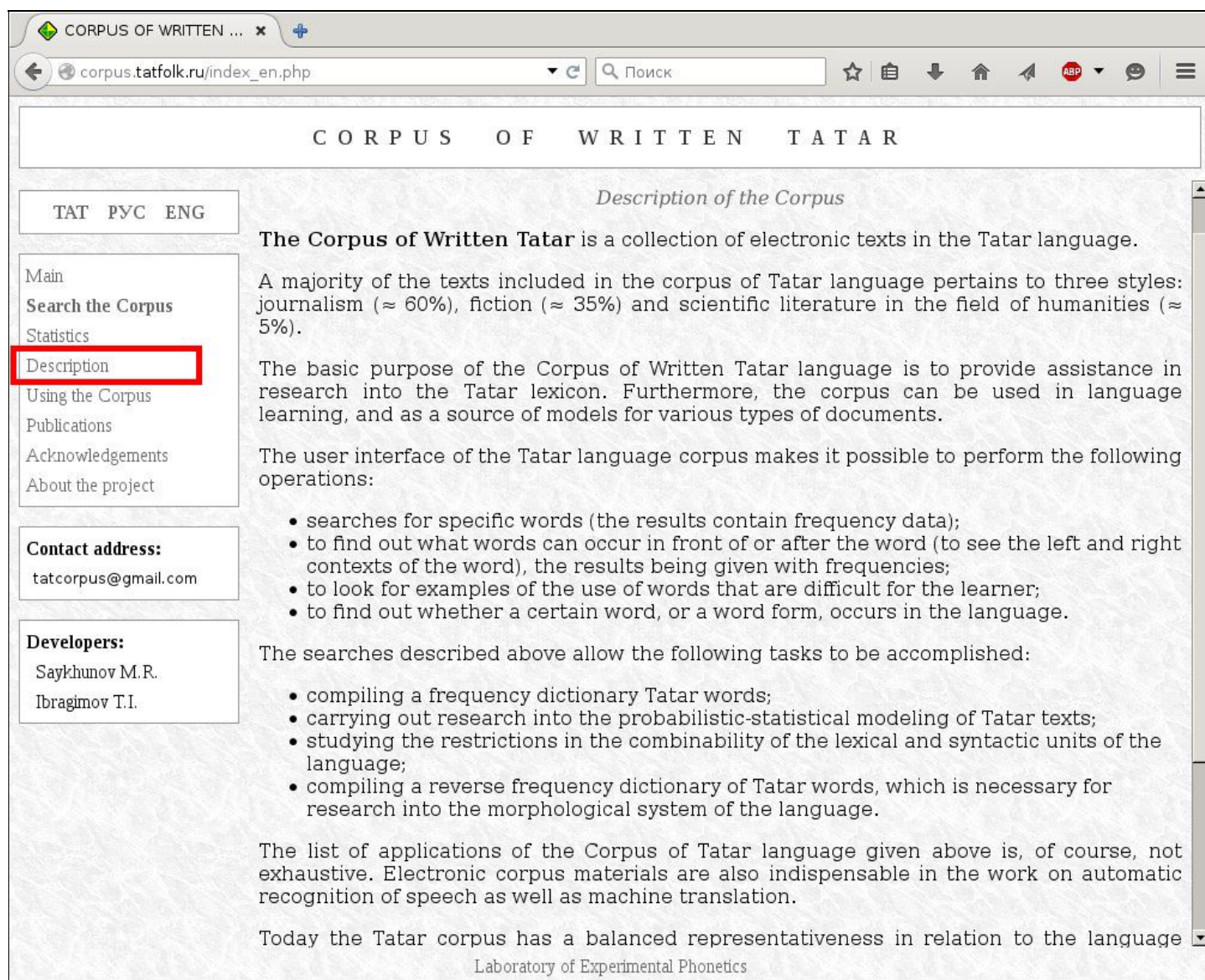
2014

Письменный корпус татарского языка: идеи, проблемы, решения // Нематериальное культурное наследие тюркских народов как объект сохранения: сборник материалов Международной научно-практической конференции (16-19 июля 2014 г.)
Письменный корпус татарского языка с озвучением визуализированных предложений как инструмент лингвистических исследований // Сопоставительная филология и полилингвизм: Материалы Всероссийской научно-практической конференции (Казань, 19-21 ноября 2014 г.)
Письменный корпус татарского языка: структурные и функциональные характеристики

Laboratory of Experimental Phonetics

Description

The website's "**Description**" section provides brief information on the structure and functionalities of the Corpus of Written Tatar:



The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". The main content area is titled "Description of the Corpus".

TAT PYC ENG

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
tatcorpus@gmail.com

Developers:
Saykhunov M.R.
Ibragimov T.I.

The Corpus of Written Tatar is a collection of electronic texts in the Tatar language.

A majority of the texts included in the corpus of Tatar language pertains to three styles: journalism ($\approx 60\%$), fiction ($\approx 35\%$) and scientific literature in the field of humanities ($\approx 5\%$).

The basic purpose of the Corpus of Written Tatar language is to provide assistance in research into the Tatar lexicon. Furthermore, the corpus can be used in language learning, and as a source of models for various types of documents.

The user interface of the Tatar language corpus makes it possible to perform the following operations:

- searches for specific words (the results contain frequency data);
- to find out what words can occur in front of or after the word (to see the left and right contexts of the word), the results being given with frequencies;
- to look for examples of the use of words that are difficult for the learner;
- to find out whether a certain word, or a word form, occurs in the language.

The searches described above allow the following tasks to be accomplished:

- compiling a frequency dictionary Tatar words;
- carrying out research into the probabilistic-statistical modeling of Tatar texts;
- studying the restrictions in the combinability of the lexical and syntactic units of the language;
- compiling a reverse frequency dictionary of Tatar words, which is necessary for research into the morphological system of the language.

The list of applications of the Corpus of Tatar language given above is, of course, not exhaustive. Electronic corpus materials are also indispensable in the work on automatic recognition of speech as well as machine translation.

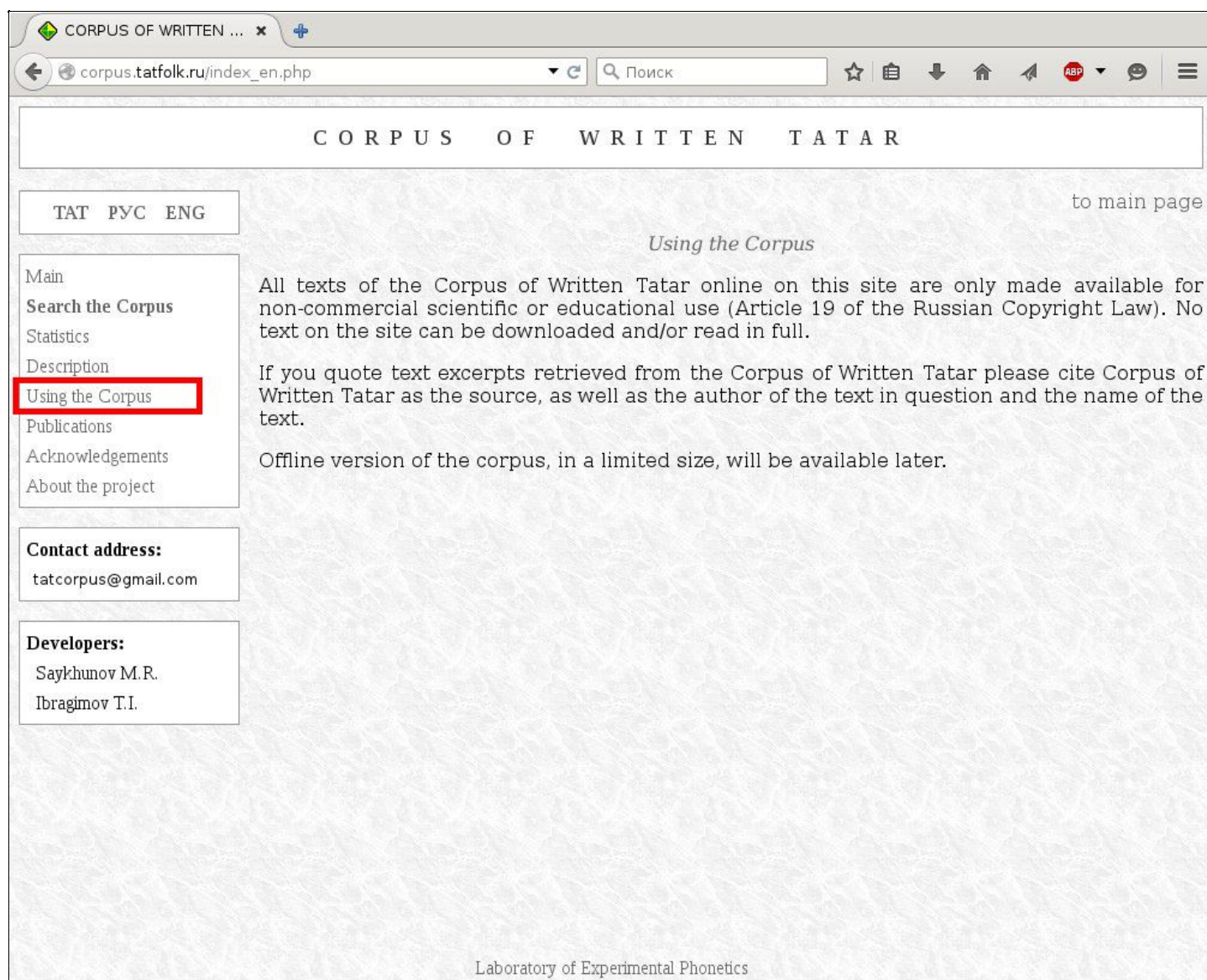
Today the Tatar corpus has a balanced representativeness in relation to the language

Laboratory of Experimental Phonetics

Use and citation

The section "**Using the Corpus**" informs the user on the legal regulations concerning the use of the Corpus, and it also gives recommendations for citing the texts and referencing to the materials of the Corpus, as well as the Corpus itself.

If you use The Corpus of Written Tatar in your scientific or educational work, please refer to the Corpus as the source, and also mention the author(s) and name of the text.



The screenshot shows a web browser window with the address bar displaying 'corpus.tatfolk.ru/index_en.php'. The page title is 'CORPUS OF WRITTEN TATAR'. The main content area is titled 'Using the Corpus' and contains the following text:

All texts of the Corpus of Written Tatar online on this site are only made available for non-commercial scientific or educational use (Article 19 of the Russian Copyright Law). No text on the site can be downloaded and/or read in full.

If you quote text excerpts retrieved from the Corpus of Written Tatar please cite Corpus of Written Tatar as the source, as well as the author of the text in question and the name of the text.

Offline version of the corpus, in a limited size, will be available later.

The left sidebar contains a navigation menu with the following items: Main, Search the Corpus, Statistics, Description, Using the Corpus (highlighted with a red box), Publications, Acknowledgements, and About the project. Below the menu are sections for 'Contact address: tatcorpus@gmail.com' and 'Developers: Saykhunov M.R., Ibragimov T.I.'. The footer of the page reads 'Laboratory of Experimental Phonetics'.

About the project

The section "**About the project**" tells about the history of the Corpus, as well as its main developers:

The screenshot shows a web browser window with the address bar displaying 'corpus.tatfolk.ru/index_en.php'. The page title is 'CORPUS OF WRITTEN TATAR'. The main content area is titled 'About the project' and contains the following text:

The work on the **Corpus of Tatar** texts was started in 2010. The beginnings of the project were connected with discussions about two directions of research, taking place in that year at the Laboratory of Applied Linguistics and Experimental Phonetics of the Kazan Federal University (KFU):

- the development of software for machine translation (MT) of Tatar texts into one of its kindred languages, and from this language back into Tatar;
- the creation of a system for automatic recognition of Tatar speech within a restricted semantic domain.

By studying the relevant literature we became aware that modern systems of MT and automatic recognition of speech rely on national corpora of the languages in question, applying the "hypothesis — check" method. This fact urged us to commit ourselves to the creation of a similar corpus of the Tatar language.

The **Corpus of Written Tatar** is mainly based on materials available in the web. Following the web addresses given after the examples (sentences) in the search results, the user can obtain more information about the sites used in creating the corpus.

The texts originating from different sources have been processed before including them in the Corpus of Tatar language: html-tags have been deleted, sentences in foreign languages have been removed, the encoding of the texts has been converted into utf-8, the sentence borders have been automatically added to the material, etc.

The work on collecting materials and processing them is going on. After having learned about the existence of the Corpus of written Tatar, many writers and scholars have provided us with electronic versions of their books and articles. According to our practice, we update the published version of the Tatar corpus when the word count of newly acquired contributions reaches 5-6 million word occurrences. At the same time, the user interface is updated.

The Corpus of Written Tatar can also be regarded as an enormous reference book, giving

Laboratory of Experimental Phonetics

The left sidebar contains a navigation menu with the following items: Main, Search the Corpus, Statistics, Description, Using the Corpus, Publications, Acknowledgements, and About the project (highlighted with a red box). Below the menu are sections for 'Contact address:' (tatcorpus@gmail.com) and 'Developers:' (Saykhunov M.R., Ibragimov T.I.).

Acknowledgments

In the section "**Acknowledgments**", the creators of the Corpus wish to express their appreciation to people and organizations that provided great assistance in the work:

The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". In the top right corner, there is a link "to main page". Below the title, there are language options: "TAT PYC ENG". The main content area is titled "Acknowledgements" and contains the text: "We express our gratitude and deep appreciation:". This is followed by a bulleted list of acknowledgments. On the left side, there is a navigation menu with the following items: "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements" (highlighted with a red box), and "About the project". Below the menu, there is a "Contact address:" section with the email `tatcorpus@gmail.com`, and a "Developers:" section listing "Saykhunov M.R." and "Ibragimov T.I.". At the bottom of the page, it says "Laboratory of Experimental Phonetics".

CORPUS OF WRITTEN TATAR

TAT PYC ENG [to main page](#)

Acknowledgements

We express our gratitude and deep appreciation:

- To the Republic Center of Development of Traditional Culture, especially to its director Fanzilya Hakimovna Zavgarova, for the place provided under the Center's web-hosting `www.tatfolk.ru`, as well as for the texts given for inclusion in the corpus.
- To the Department of Finno-Ugric languages at the Turku University (Finland), and personally to Jorma Luutonen, for support and valuable advices.
- To the editorial office of the popular scientific journal "Фән һәм Тел", to the editor-in-chief Rashit Agleevich Shakirzyanov, and also to Farid Shakirzyanov, for the texts and assistance provided to the project.
- To Chusainov R.R. ("GDC" company) for great help during the whole period of development of the Corpus.
- To Chugunov A.N. ("RX5" company) for the valuable advice on database design and optimization of search queries.
- TO ALL AUTHORS WHO SENT US THEIR WORKS AND WRITINGS!!!

Main
Search the Corpus
Statistics
Description
Using the Corpus
Publications
Acknowledgements
About the project

Contact address:
`tatcorpus@gmail.com`

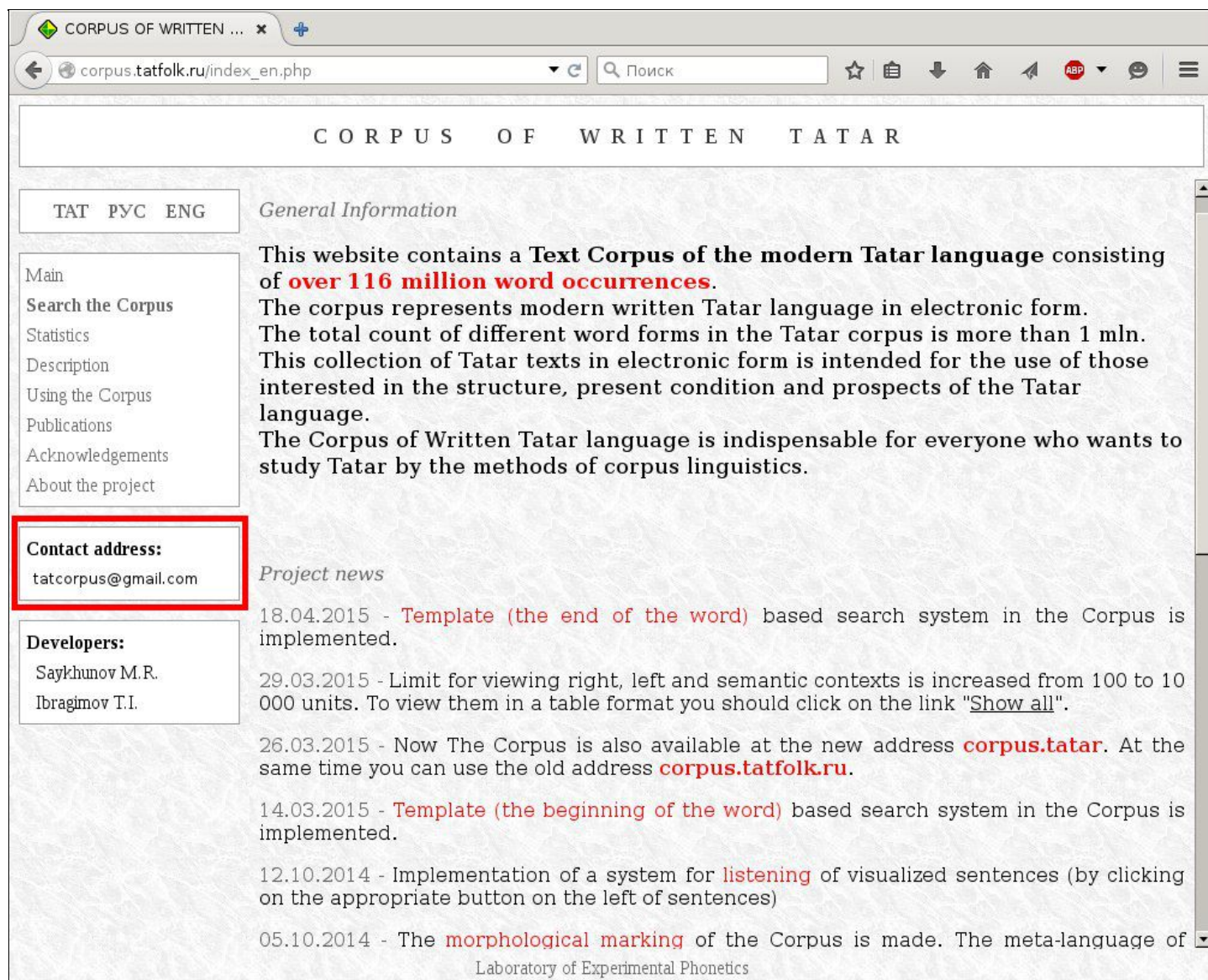
Developers:
Saykhunov M.R.
Ibragimov T.I.

Laboratory of Experimental Phonetics

Contacts

You can reach the developers of the corpus by sending a message to tatcorpus@gmail.com

Please contact us if you have any questions or suggestions concerning the Corpus and its use!

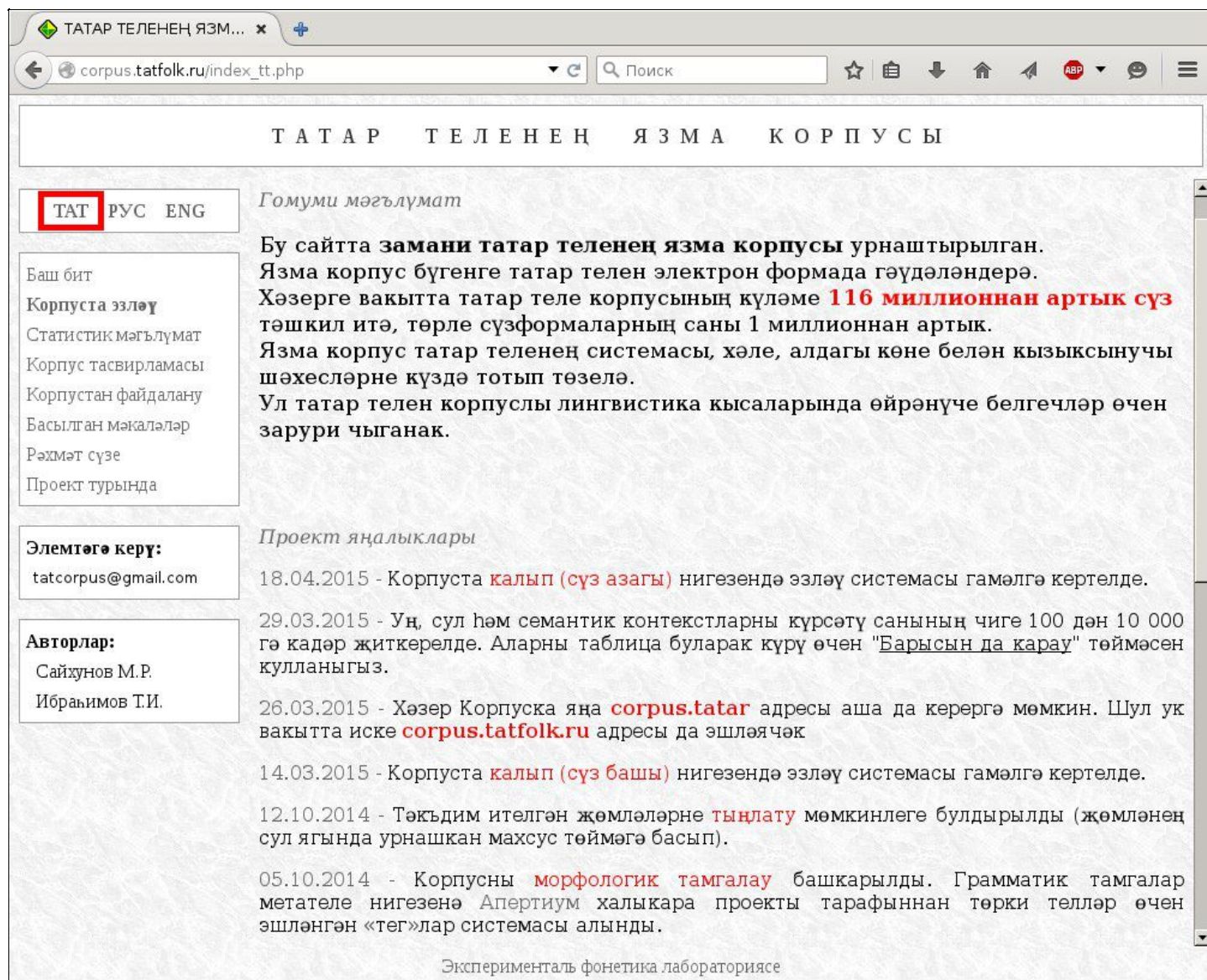


The screenshot shows a web browser window with the URL `corpus.tatfolk.ru/index_en.php`. The page title is "CORPUS OF WRITTEN TATAR". On the left, there is a navigation menu with options: "Main", "Search the Corpus", "Statistics", "Description", "Using the Corpus", "Publications", "Acknowledgements", and "About the project". Below the menu, there are two boxes: "Contact address:" containing `tatcorpus@gmail.com` (highlighted with a red border) and "Developers:" listing "Saykhunov M.R." and "Ibragimov T.I.". The main content area is titled "General Information" and contains the following text: "This website contains a **Text Corpus of the modern Tatar language** consisting of **over 116 million word occurrences**. The corpus represents modern written Tatar language in electronic form. The total count of different word forms in the Tatar corpus is more than 1 mln. This collection of Tatar texts in electronic form is intended for the use of those interested in the structure, present condition and prospects of the Tatar language. The Corpus of Written Tatar language is indispensable for everyone who wants to study Tatar by the methods of corpus linguistics." Below this is a "Project news" section with several entries: "18.04.2015 - **Template (the end of the word)** based search system in the Corpus is implemented.", "29.03.2015 - Limit for viewing right, left and semantic contexts is increased from 100 to 10 000 units. To view them in a table format you should click on the link **'Show all'**.", "26.03.2015 - Now The Corpus is also available at the new address **corpus.tatar**. At the same time you can use the old address **corpus.tatfolk.ru**.", "14.03.2015 - **Template (the beginning of the word)** based search system in the Corpus is implemented.", "12.10.2014 - Implementation of a system for **listening** of visualized sentences (by clicking on the appropriate button on the left of sentences)", and "05.10.2014 - The **morphological marking** of the Corpus is made. The meta-language of Laboratory of Experimental Phonetics".

In other languages

For the convenience of our users, all sections of the website of the Corpus of Written Tatar are available in three languages: Russian, Tatar and English. Work is underway to translate them into other languages!

To choose the interface language, click on the appropriate link above the main menu on the left of the screen (in this case, TAT for Tatar language):



The screenshot shows a web browser window with the address bar displaying 'corpus.tatfolk.ru/index_tt.php'. The page title is 'ТАТАР ТЕЛЕНЕҢ ЯЗМА КОРПУСЫ'. In the top left corner, there is a language selection menu with three options: 'TAT', 'РУС', and 'ENG'. The 'TAT' option is highlighted with a red box. Below the menu, there are several sections: 'Баш бит' (Home), 'Корпуста эзләү' (Search in the corpus), 'Статистик мәгълүмат' (Statistics), 'Корпус тасвирламасы' (Corpus description), 'Корпустаң файдалану' (Using the corpus), 'Басылган мақалалар' (Published articles), 'Рәхмәт сүзе' (Words of gratitude), and 'Проект турында' (About the project). The main content area is titled 'Гомуми мәгълүмат' (General information) and contains a paragraph in Tatar language. Below this, there is a section titled 'Проект яңалыклары' (Project news) with several dates and updates.

ТАТАР ТЕЛЕНЕҢ ЯЗМА КОРПУСЫ

TAT РУС ENG

Баш бит
Корпуста эзләү
Статистик мәгълүмат
Корпус тасвирламасы
Корпустаң файдалану
Басылган мақалалар
Рәхмәт сүзе
Проект турында

Элемтәгә керү:
tatcorpus@gmail.com

Авторлар:
Сайхунов М.Р.
Ибраһимов Т.И.

Гомуми мәгълүмат

Бу сайтта замани татар теленең язма корпусы урнаштырылган. Язма корпус бүгенге татар телен электрон формада гәүдәләндерә. Хәзерге вакытта татар теле корпусының күләме **116 миллионнан артык сүз** тәшкил итә, төрле сүзформаларның саны 1 миллионнан артык. Язма корпус татар теленең системасы, хәле, алдагы көне белән кызыксынучы шәхесләрне күздә тотып төзелә. Ул татар телен корпуслы лингвистика кысаларында өйрәнүче белгечләр өчен зарури чыганак.

Проект яңалыклары

18.04.2015 - Корпуста **калып (сүз азагы)** нигезендә эзләү системасы гамәлгә кертелде.

29.03.2015 - Уң, сул һәм семантик контекстларны күрсәтү санының чиге 100 дән 10 000 гә кадәр житкерелде. Аларны таблица буларак күрү өчен "Барысын да карау" төймәсен кулланыгыз.

26.03.2015 - Хәзер Корпуска яңа **corpus.tatar** адресы аша да керергә мөмкин. Шул ук вакытта искә **corpus.tatfolk.ru** адресы да эшләчәк

14.03.2015 - Корпуста **калып (сүз башы)** нигезендә эзләү системасы гамәлгә кертелде.

12.10.2014 - Тәкъдим ителгән **жөмлөләрне тыңлату** мөмкинлегә булдырылды (жөмләнән сул ягында урнашкан махсус төймәгә басып).

05.10.2014 - Корпусны **морфологик тамгалау** башкарылды. Грамматик тамгалар метателе нигезенә Апертиум халыкара проекты тарафыннан төрки телләр өчен эшләнгән «тег»лар системасы алынды.

Эксперименталь фонетика лабораториясе

Below, the Russian interface of the Corpus website has been chosen:

ПИСЬМЕННЫЙ КОРПУС ТАТАРСКОГО ЯЗЫКА

TAT **РУС** ENG

Общая информация

На этом сайте помещен **Письменный корпус современного татарского языка**. Корпус представляет современный письменный татарский язык в электронной форме. Объем корпуса татарского языка в настоящее время составляет **более 116 млн. слов**, число различных словоформ – более 1 млн. Электронный корпус предназначен интересующимся системой, состоянием и перспективой татарского языка. Он необходим лингвистам, изучающим татарский язык в рамках корпусной лингвистики.

Новости проекта

18.04.2015 - Внедрена система поиска в Корпусе **по шаблону (конец слова)**.

29.03.2015 - Лимит на просмотр правого, левого и семантического контекстов увеличен со 100 до 10 000 единиц. Для их просмотра в табличном виде необходимо нажать ссылку "[Показать все](#)".

26.03.2015 - Теперь Корпус доступен и по новому адресу **corpus.tatar**. Доступ по старому адресу **corpus.tatfolk.ru** сохранен.

14.03.2015 - Внедрена система поиска в Корпусе **по шаблону (начало слова)**.

12.10.2014 - Реализована возможность **прослушивания** визуализированных предложений (нажав на соответствующую кнопку слева от предложения).

05.10.2014 - Произведена **морфологическая разметка** Корпуса. В основу метаязыка

Лаборатория экспериментальной фонетики