CORPUS OF WRITTEN TATAR

User's Guide

v3.4

Mansur Saykhunov

Tavzikh Ibragimov

Jorma Luutonen

Contents

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 (≈ 60% of the total volume of the Corpus) represents the publicistic style. Other styles covered by the Corpus are the following:

- fiction ≈ 35%,
- scientific literature (Humanities) ≈ 4%,
- official and business papers ≈ 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:

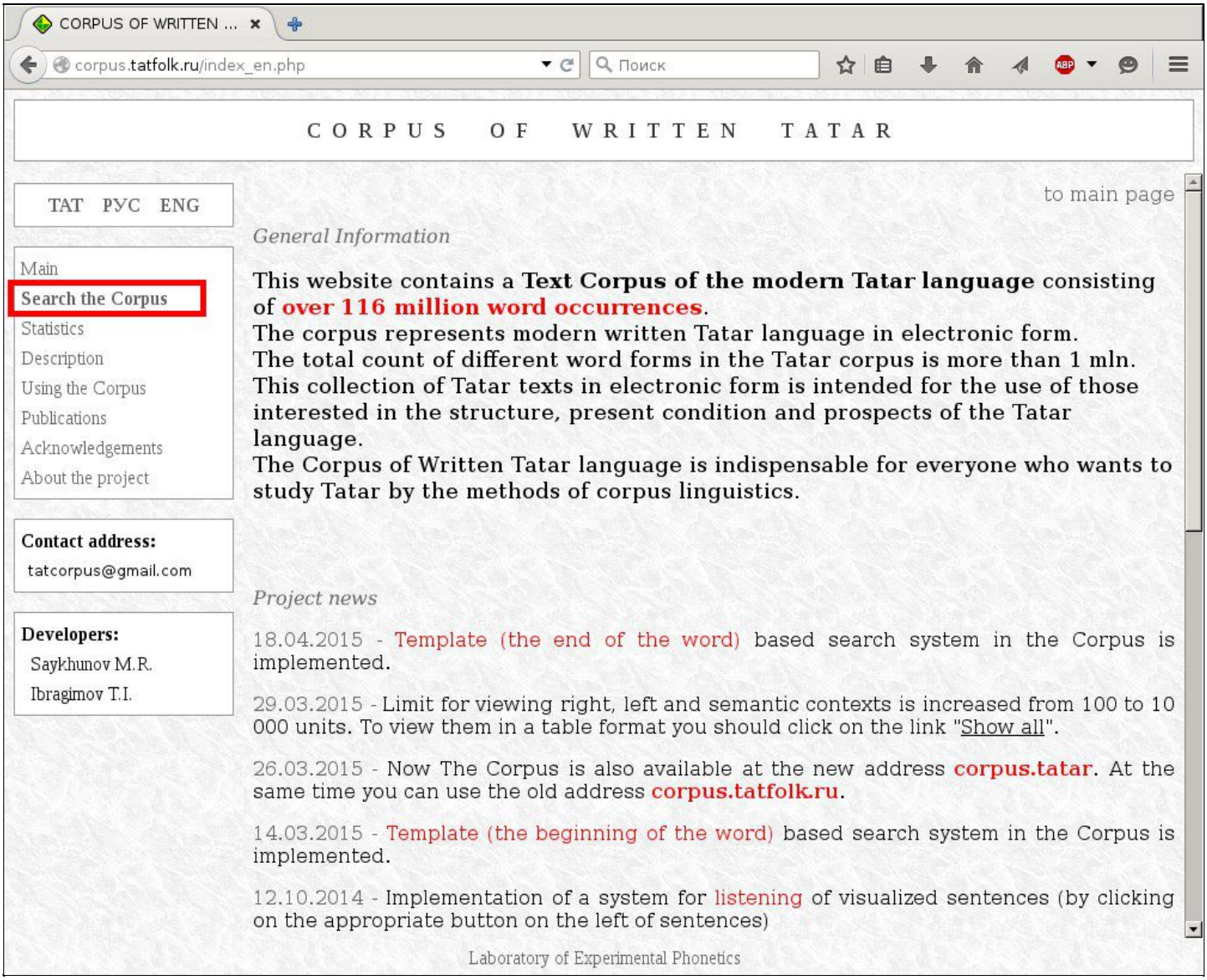
1. Use the most recent versions of web-browsers:
    - Mozilla Firefox: https://www.mozilla.org/en-US/firefox/all/
    - Google Chrome: https://www.google.ru/intl/en/chrome/browser/desktop/index.html

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

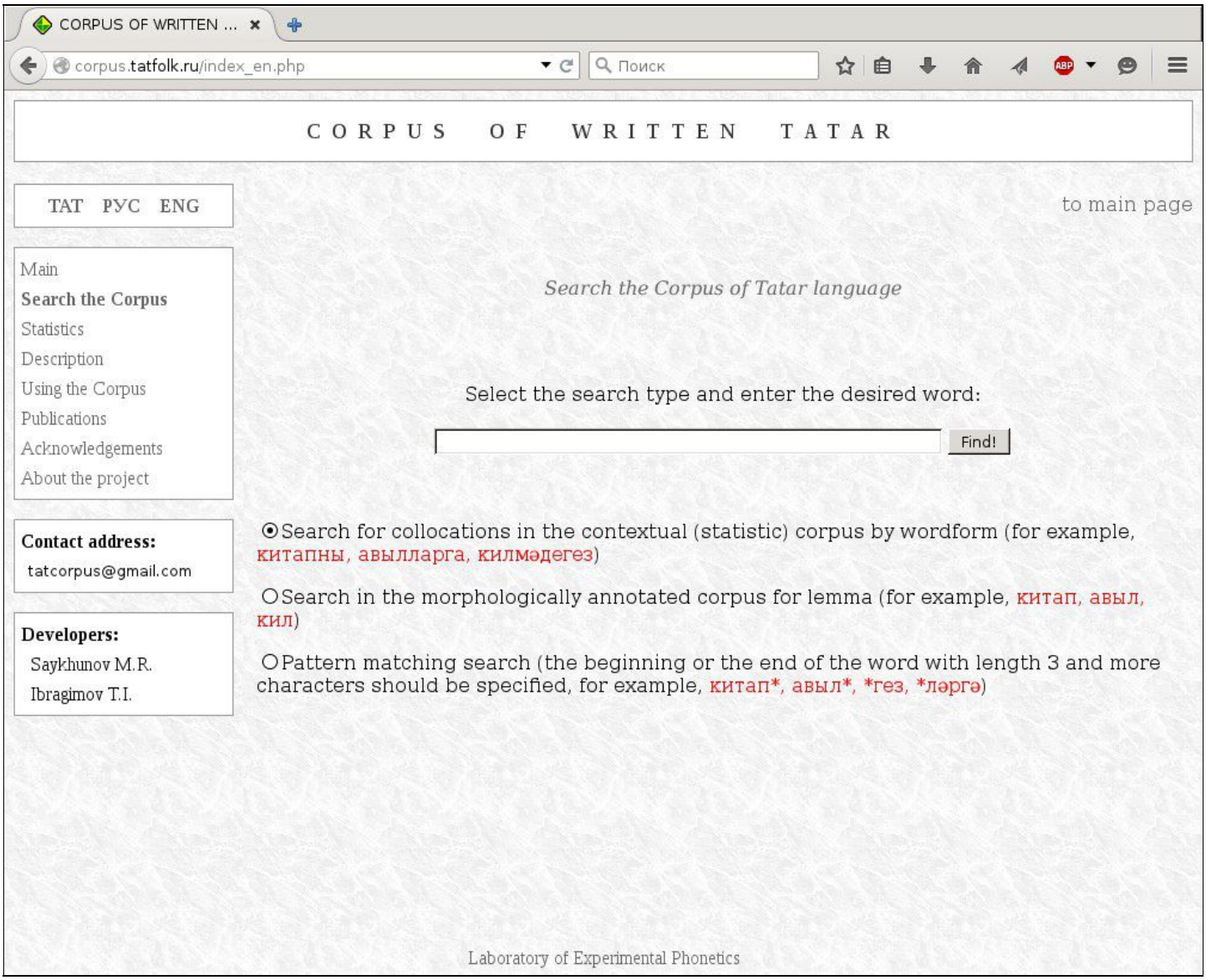
3. If you have problems with listening to the sentences:
    - Try to use another web-browser
    - Install new version of Adobe Flash Player (https://get.adobe.com/flashplayer/)

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:

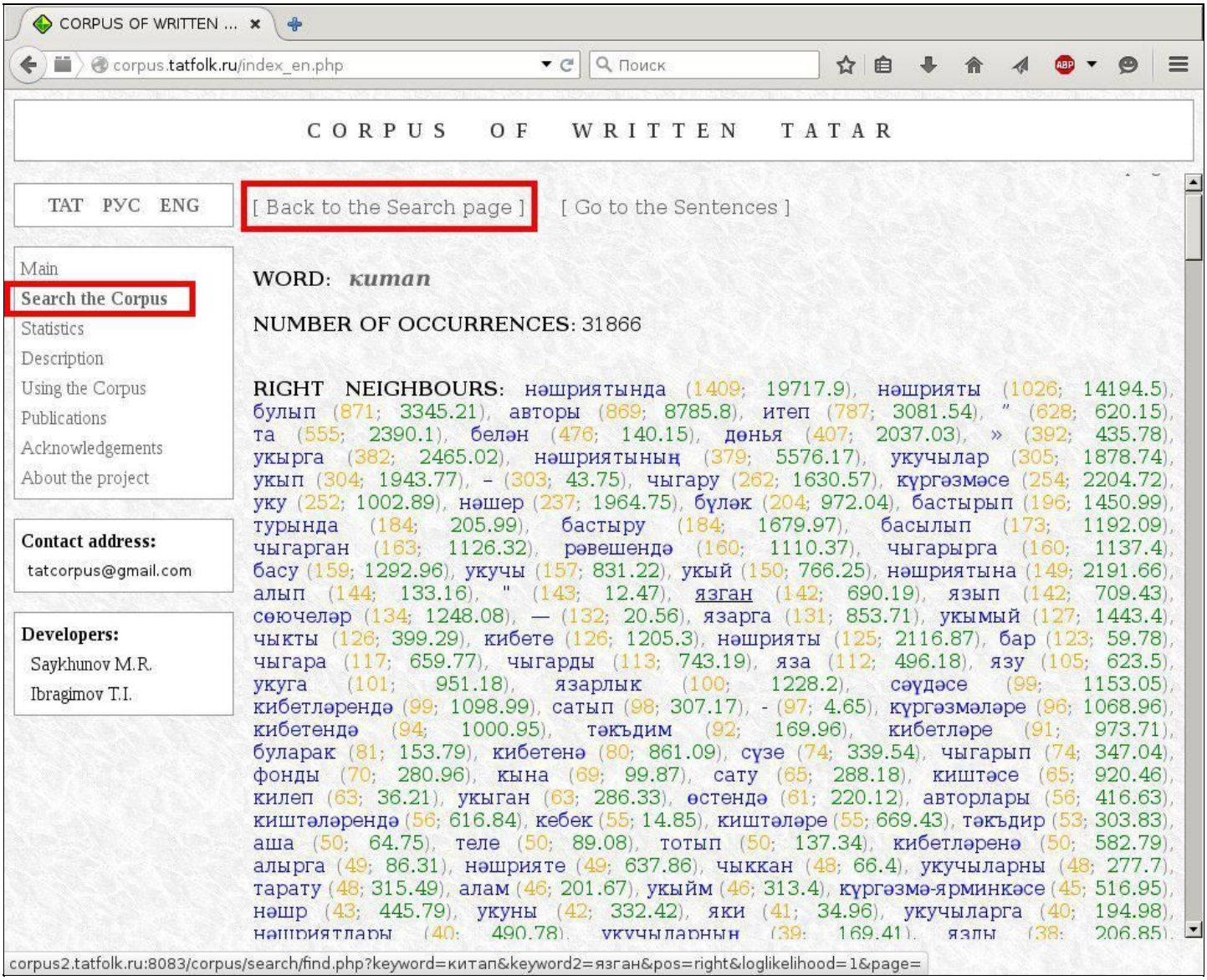


As a result, the "**Search in the Corpus of Tatar language**" page will open:

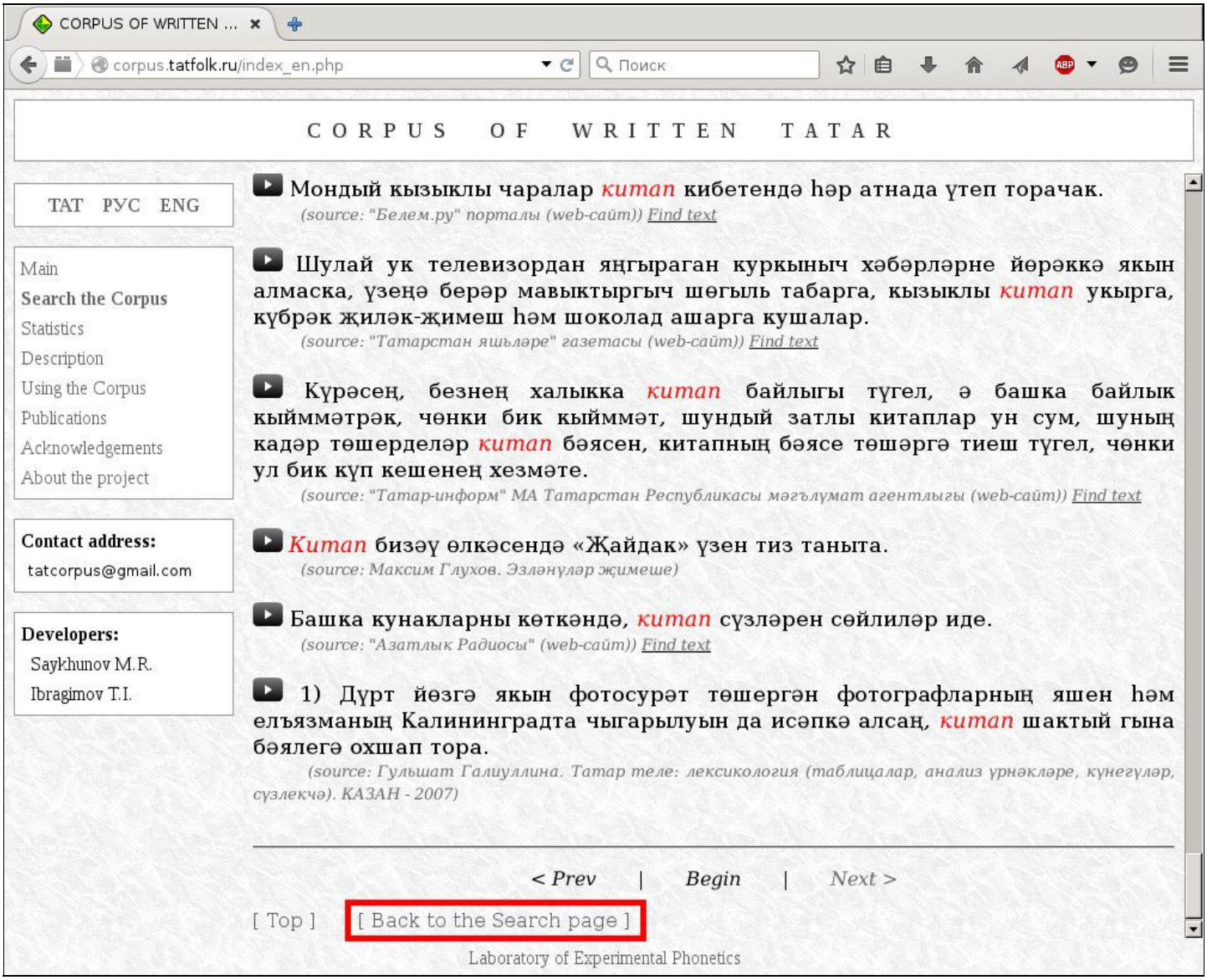


How to get back to the search page?

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



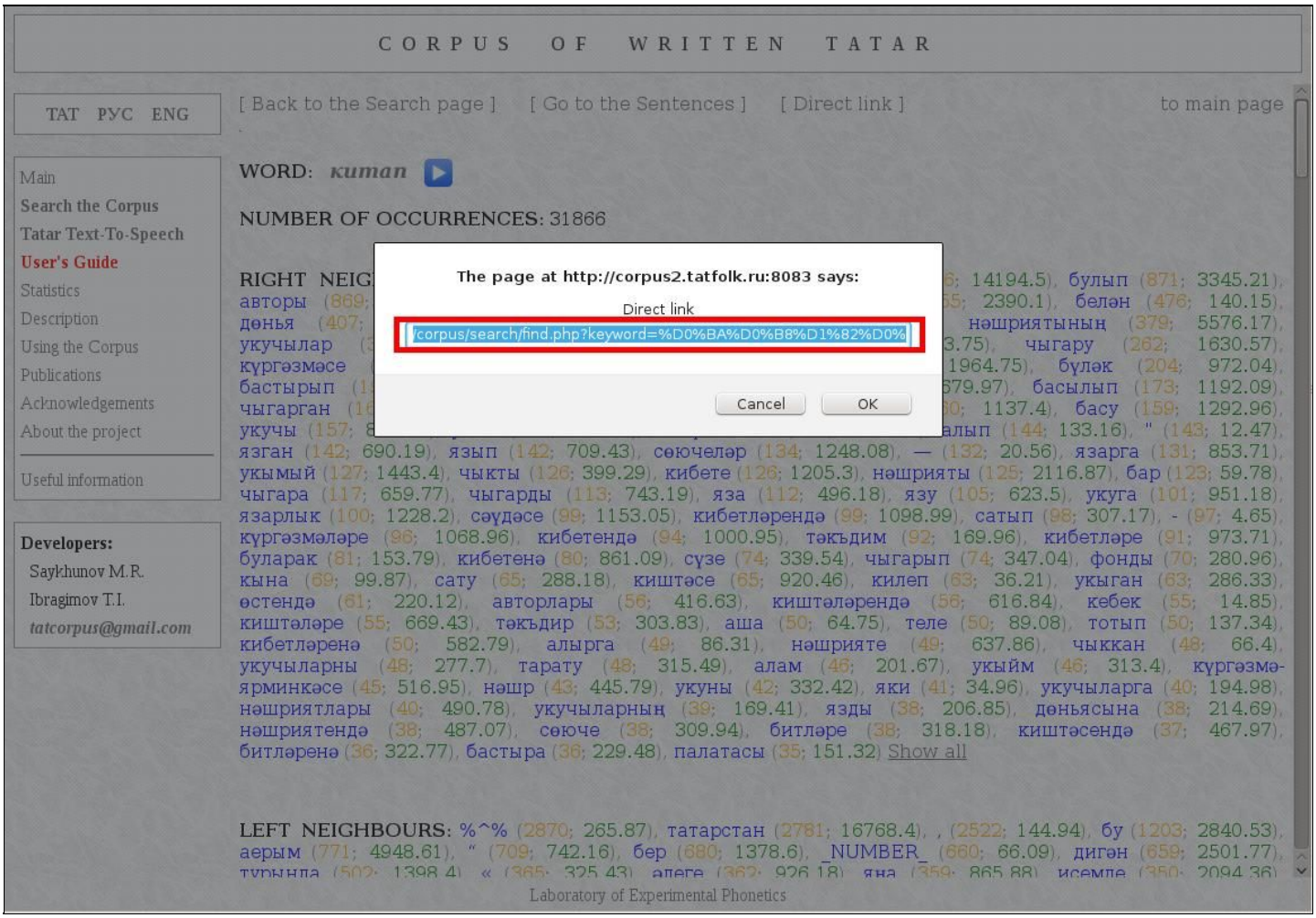
There is also a link at the bottom of the page:



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

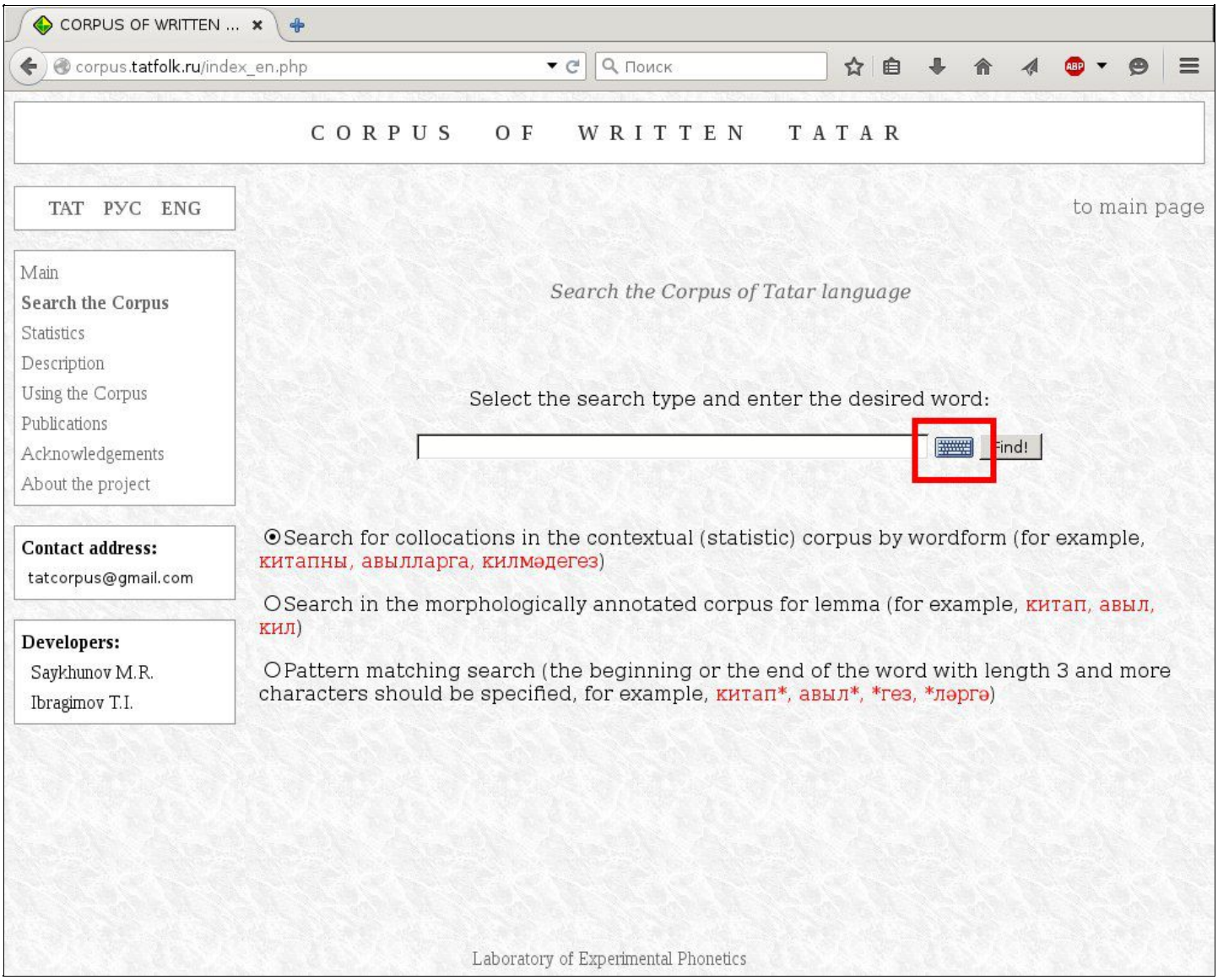
Please do not try to reference the search results with the url in the browser's address bar. Since the Corpus site has a frame-based structure, url is just a link to the home page. To get an appropriate link, you can use the "Direct Link" button at the top of the search results page:

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

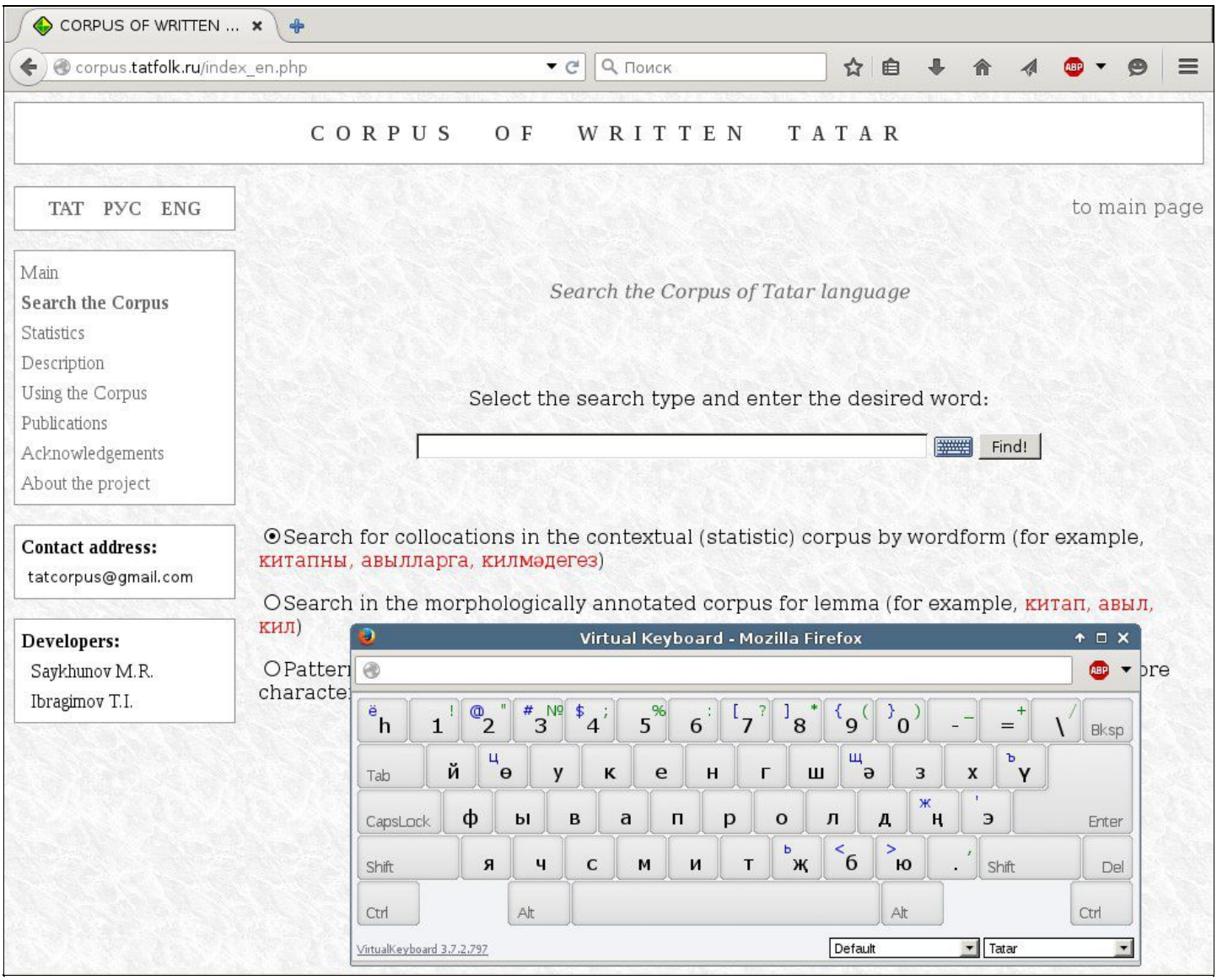


What to do if you don’t have a Tatar keyboard layout?

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.

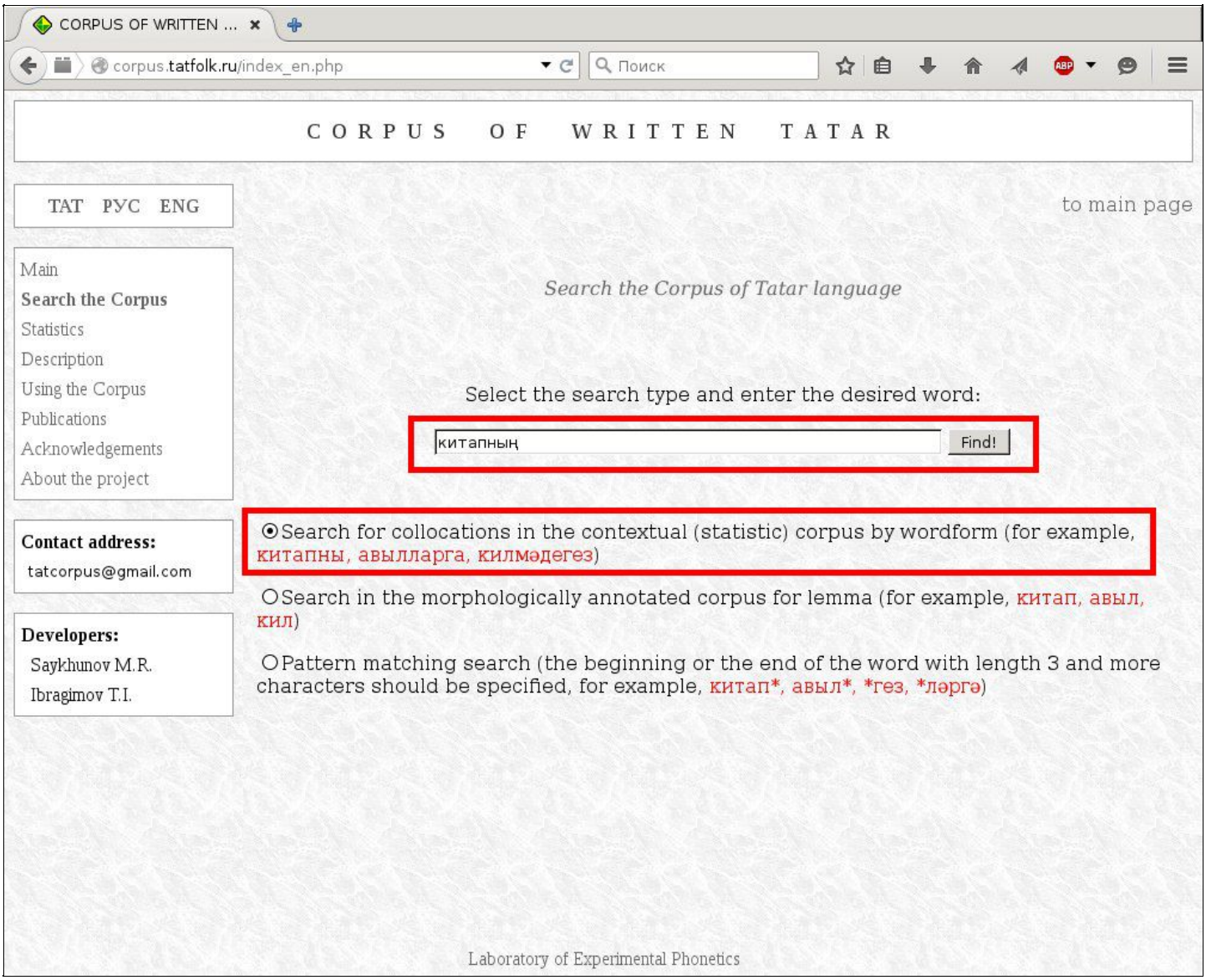


Using the virtual keyboard is very easy. When it pops up, you just write the word you want to find by clicking the letters in the virtual keyboard, or, alternatively, by pressing the corresponding buttons on your own physical keyboard.



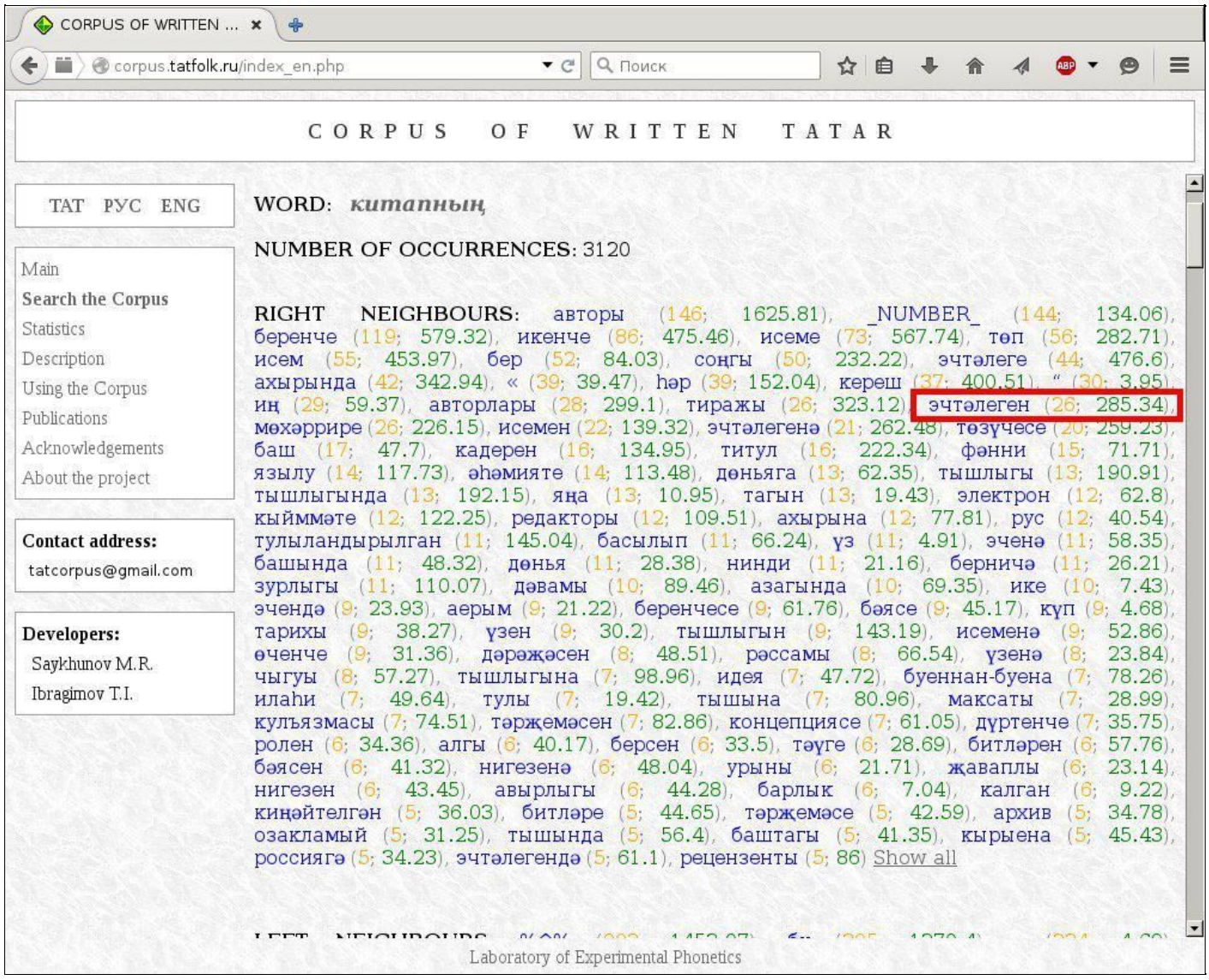
How to find a word form and reveal its contextual statistics in the Corpus?

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:



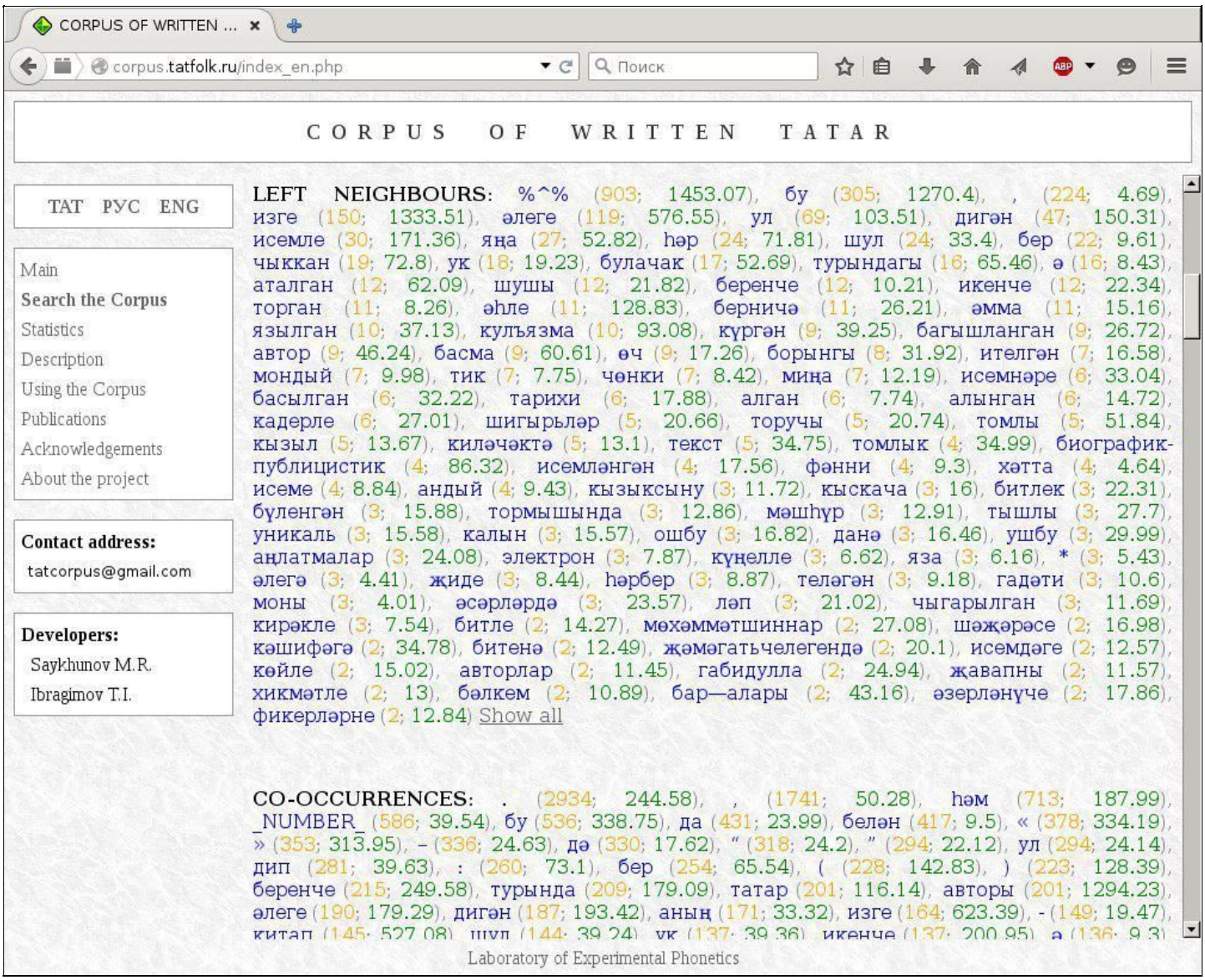
Search results will open in a new window. First, the targeted word and the number of its occurrences in the Corpus will be shown. Below these, further data is given in three sections:

- **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

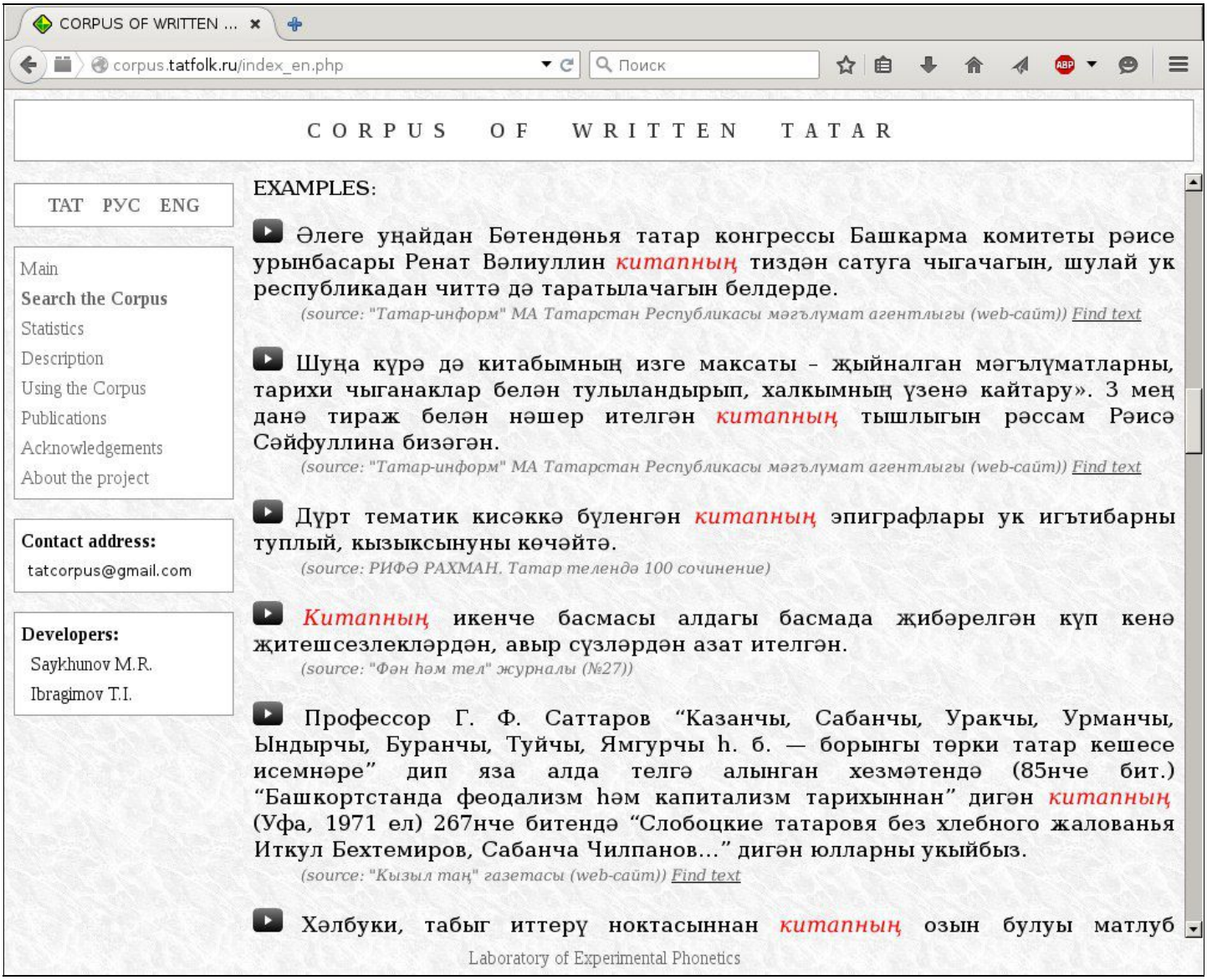


- **Left neighbours** - the same as in the preceding paragraph but applied to the left-side context of the word, i.e. what word is found immediately before the targeted word!

- **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.



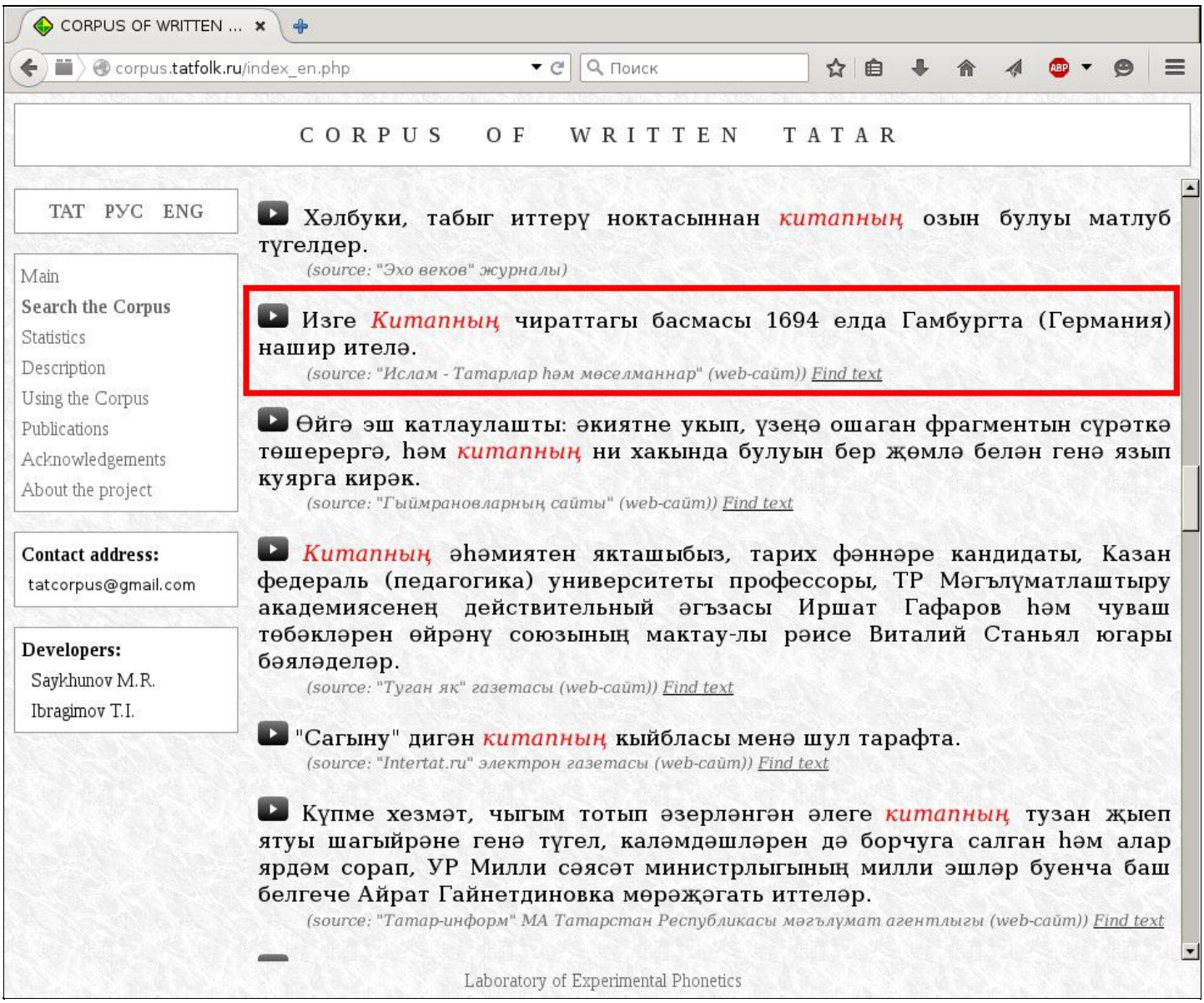
Below the co-occurrences data, examples are given:



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

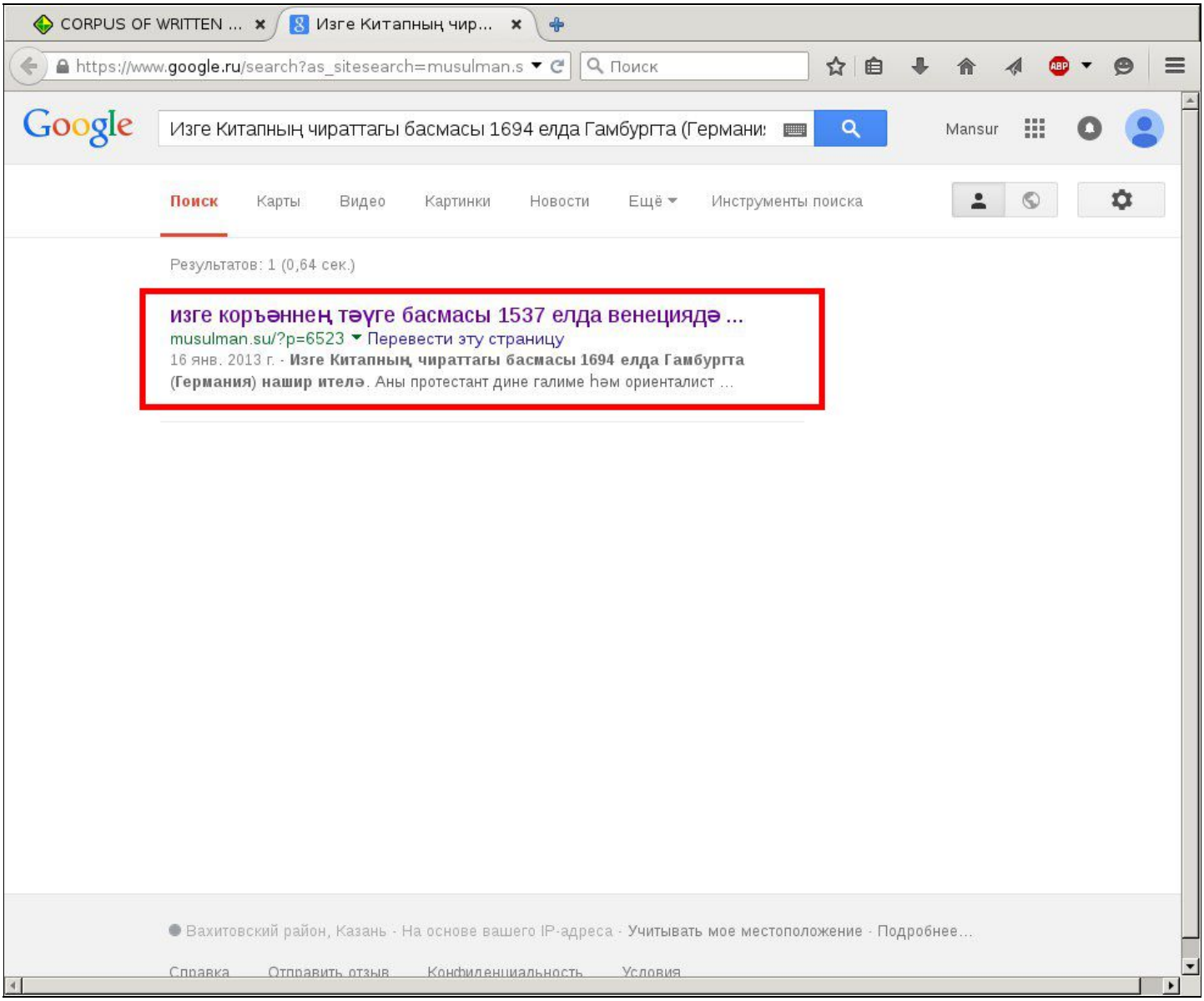
The text in small print after each sentence indicates the source from which the example was taken.

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

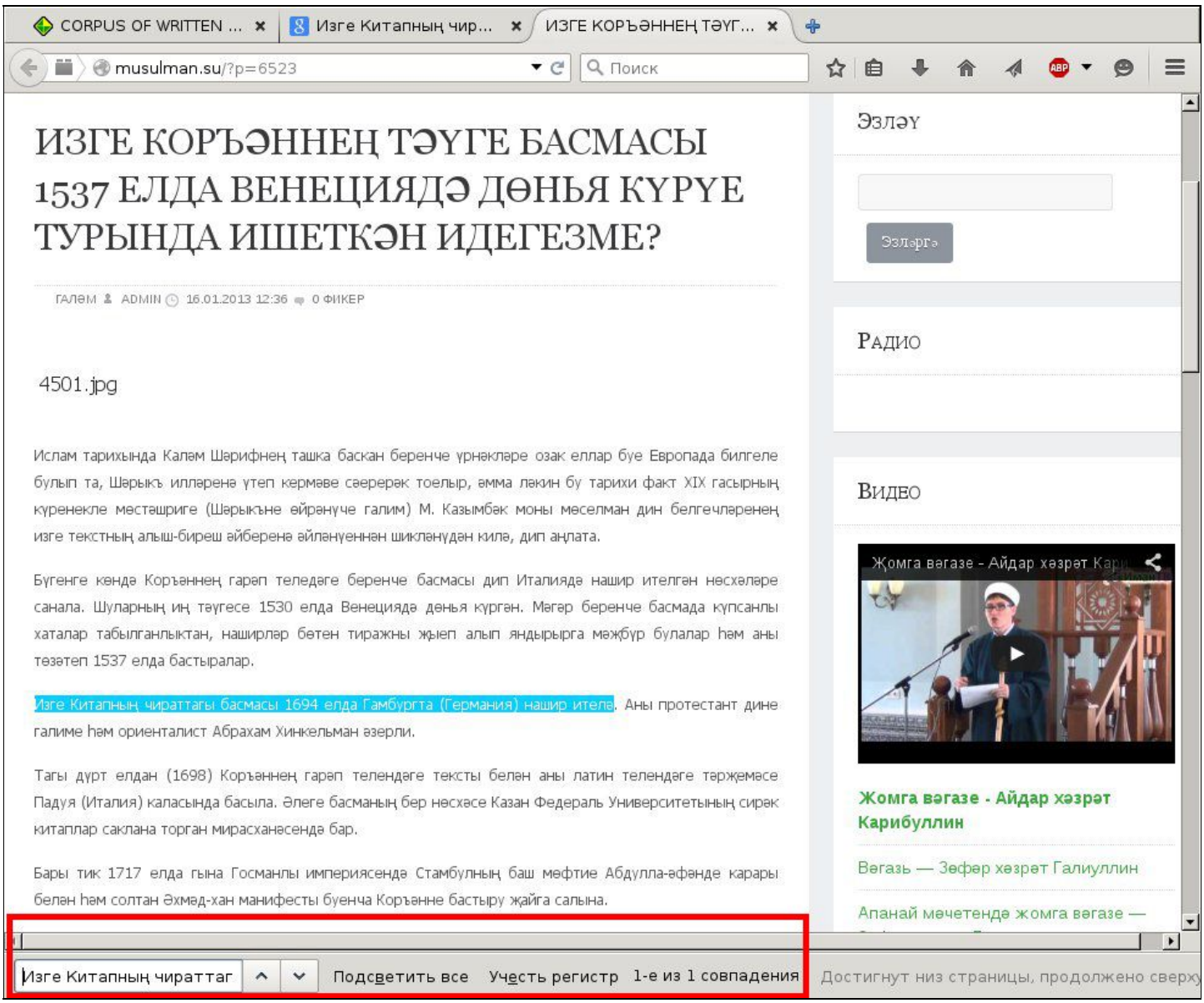


In the case of belletristic texts, there is no "**Find text**" link because of restrictions related to copyright.

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

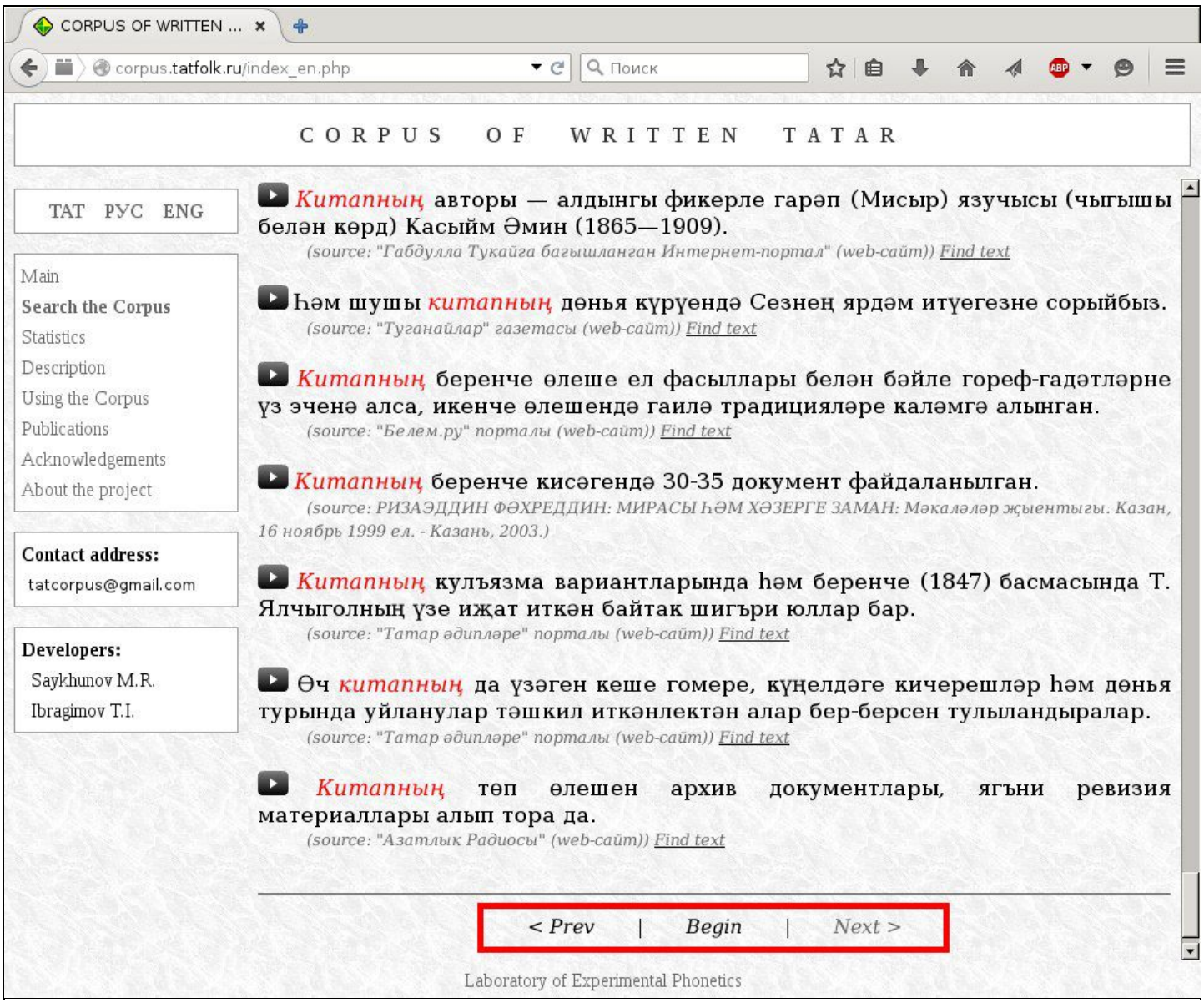


And, in most cases, the first link given by Google will refer to the page that is associated with the text containing the sentence in question:



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.

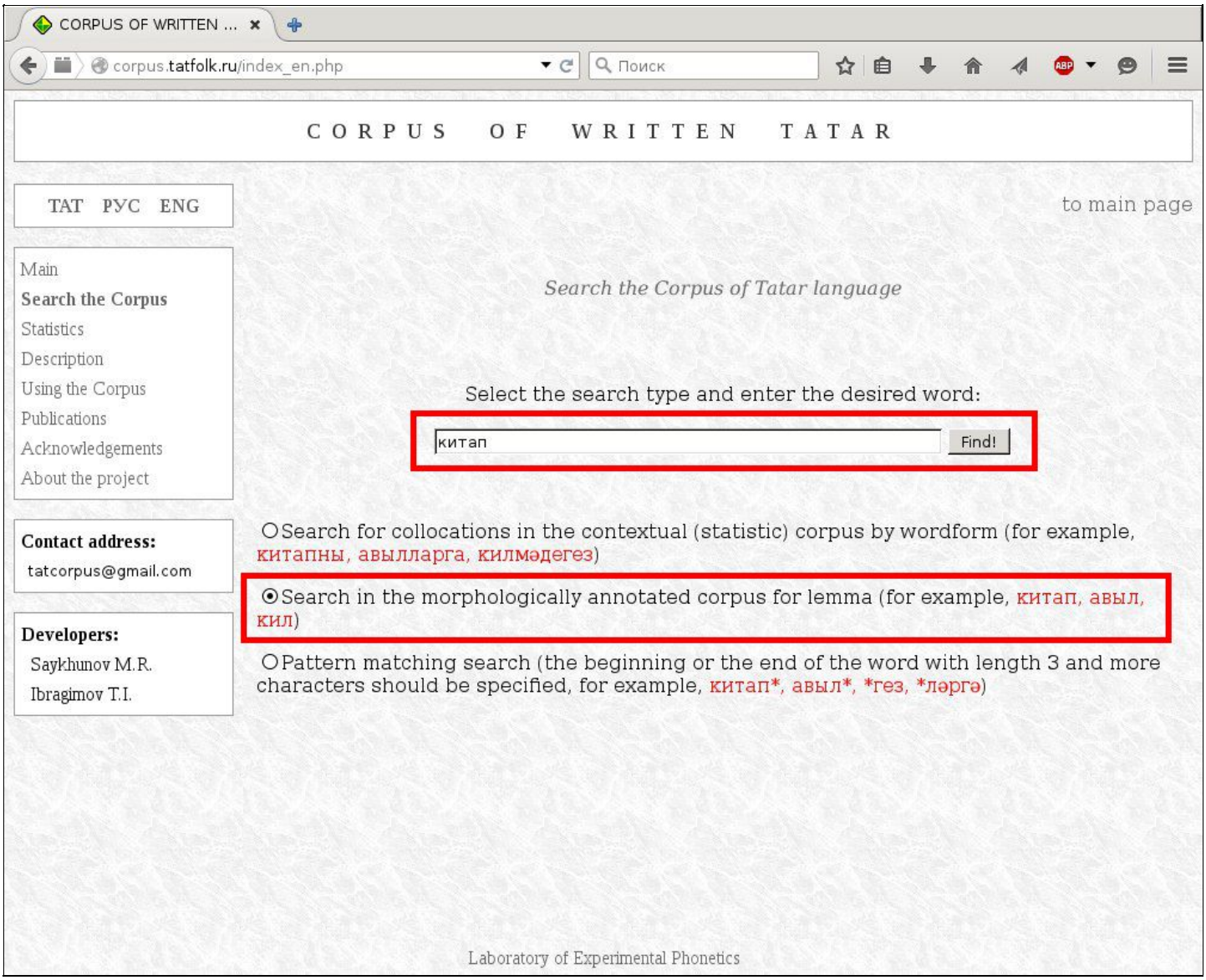
Below the list of example sentences, there are links to other pages, containing more sentences. By clicking "**Next**" the user will be able to see the following 50 examples, and so on, until all sentences containing the targeted word have been displayed.



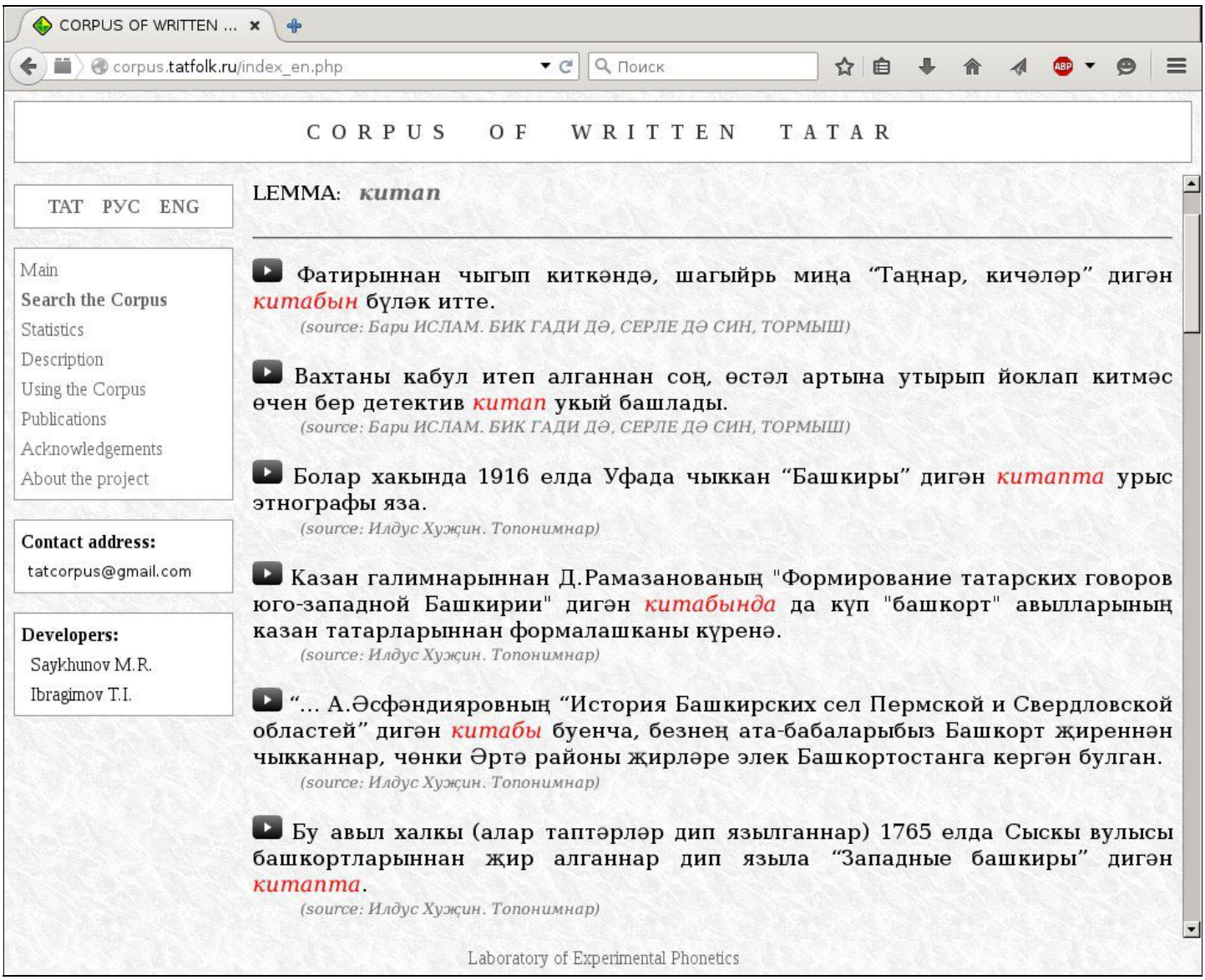
How to find a lexeme in the Corpus?

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...**".



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



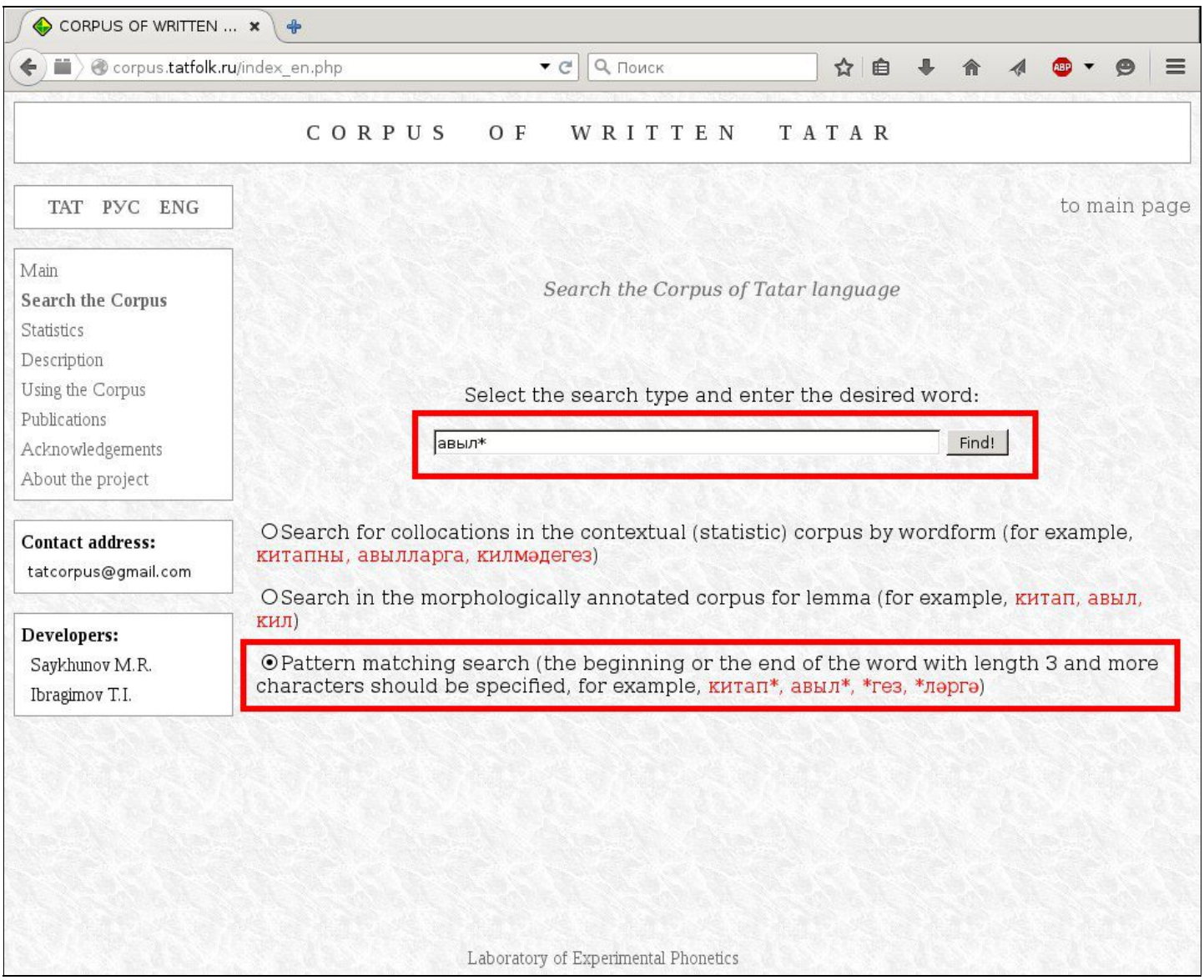
The system of source indication, as well as the full text viewing possibility, is identical with that of the basic word searches described previously.

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

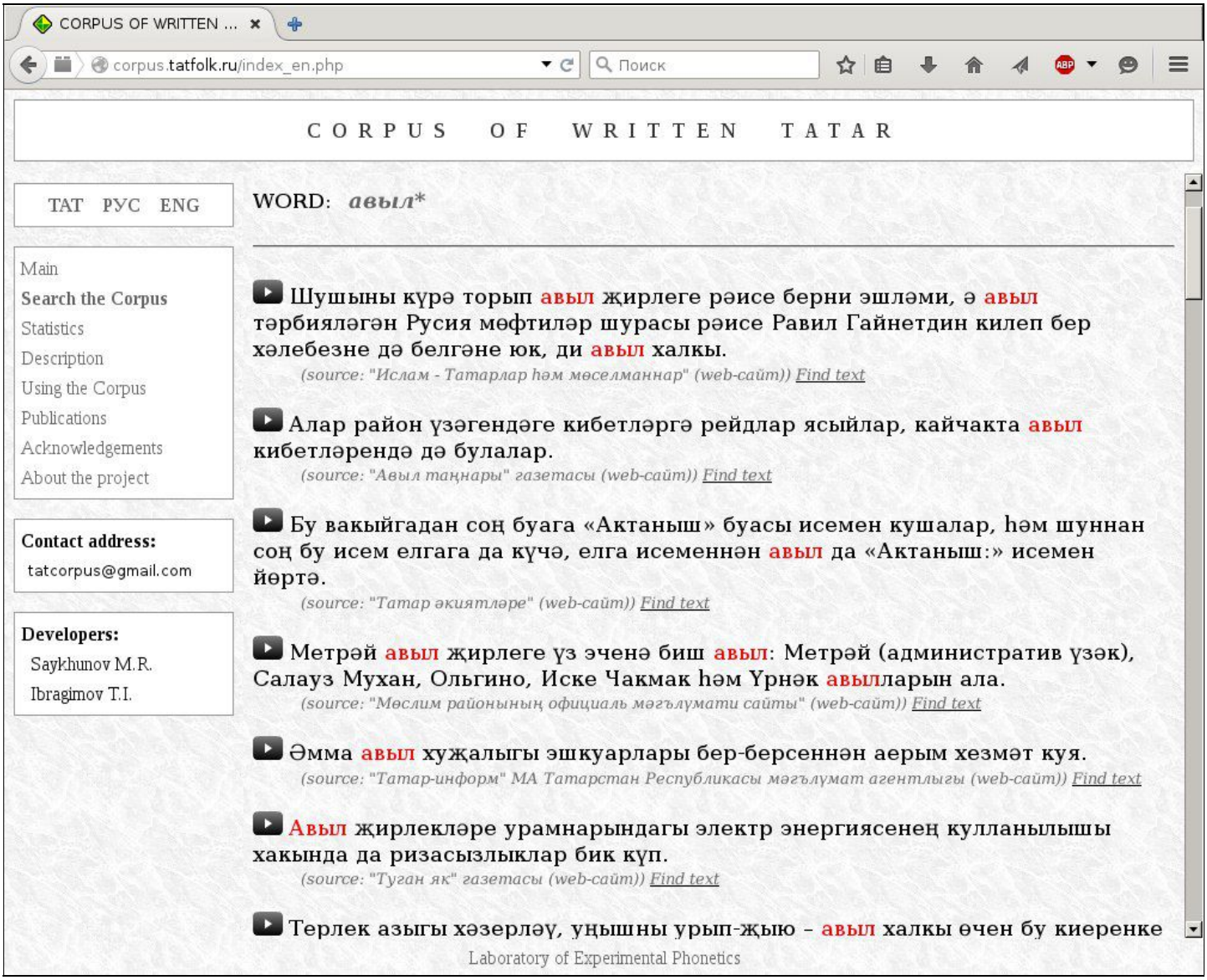
It is possible to search for lexemes having a particular initial part in the Corpus.

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.



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

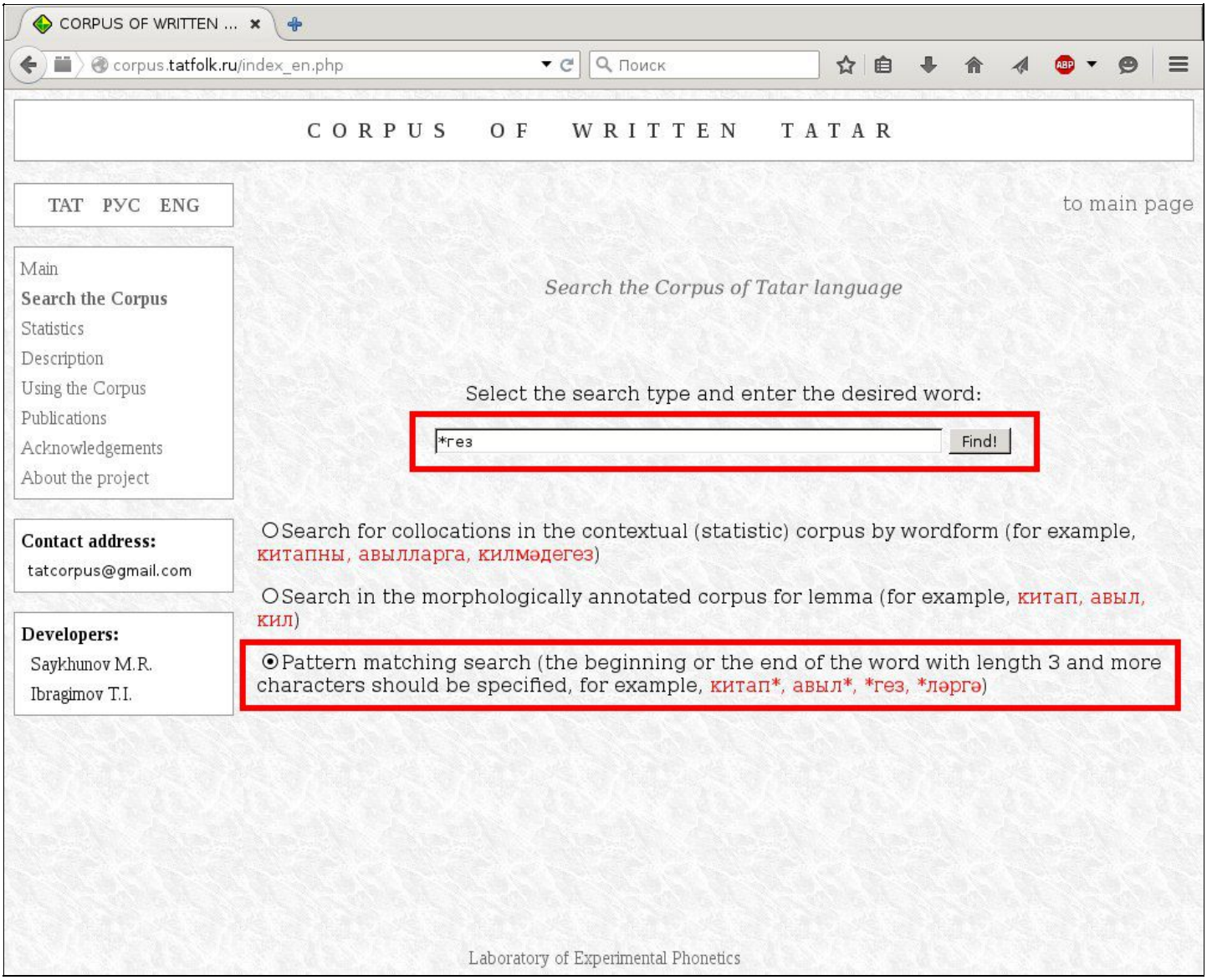


How to find all lexemes ending in a given postfix in the Corpus?

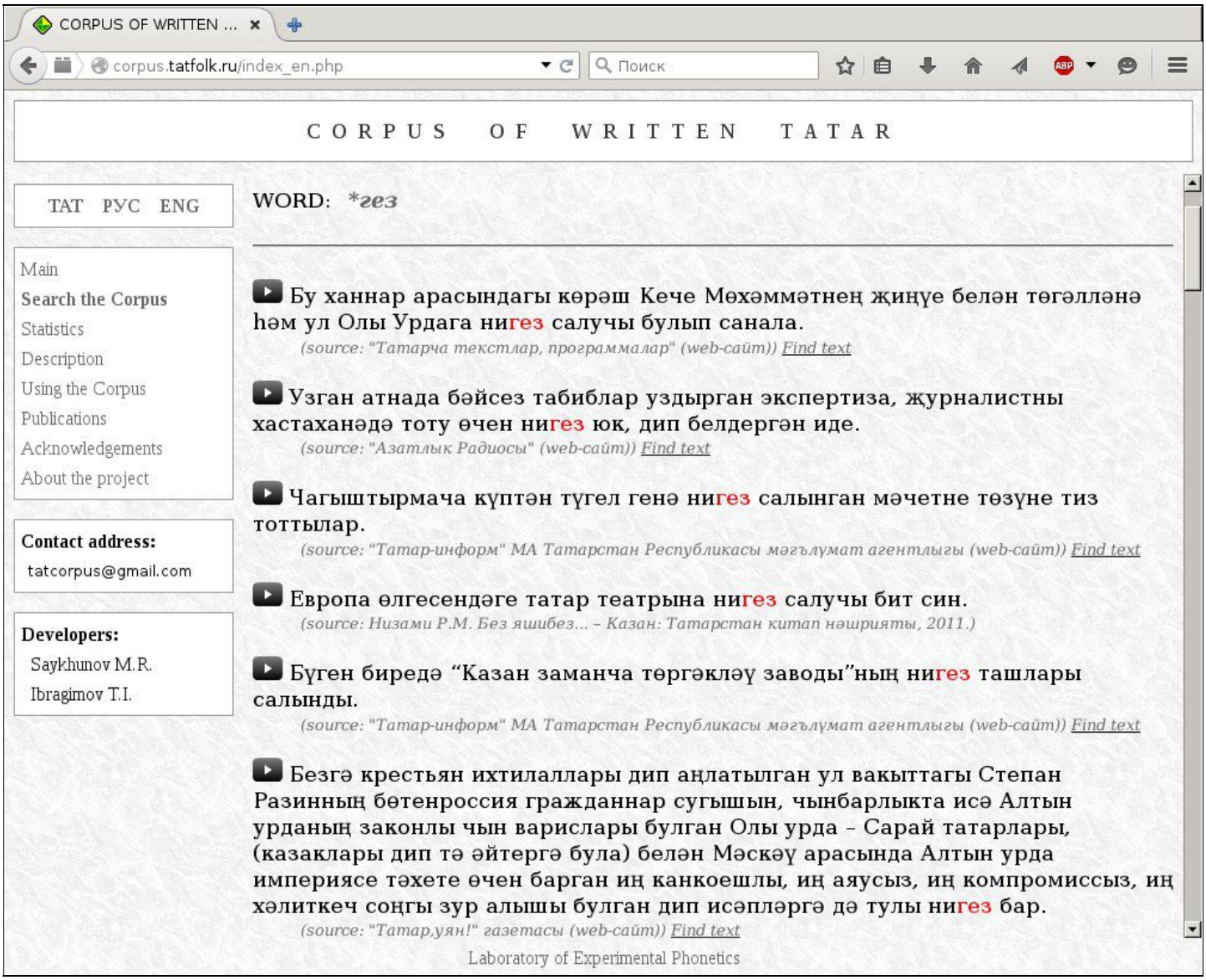
It is possible to search for lexemes having a particular final part in the Corpus.

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.



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

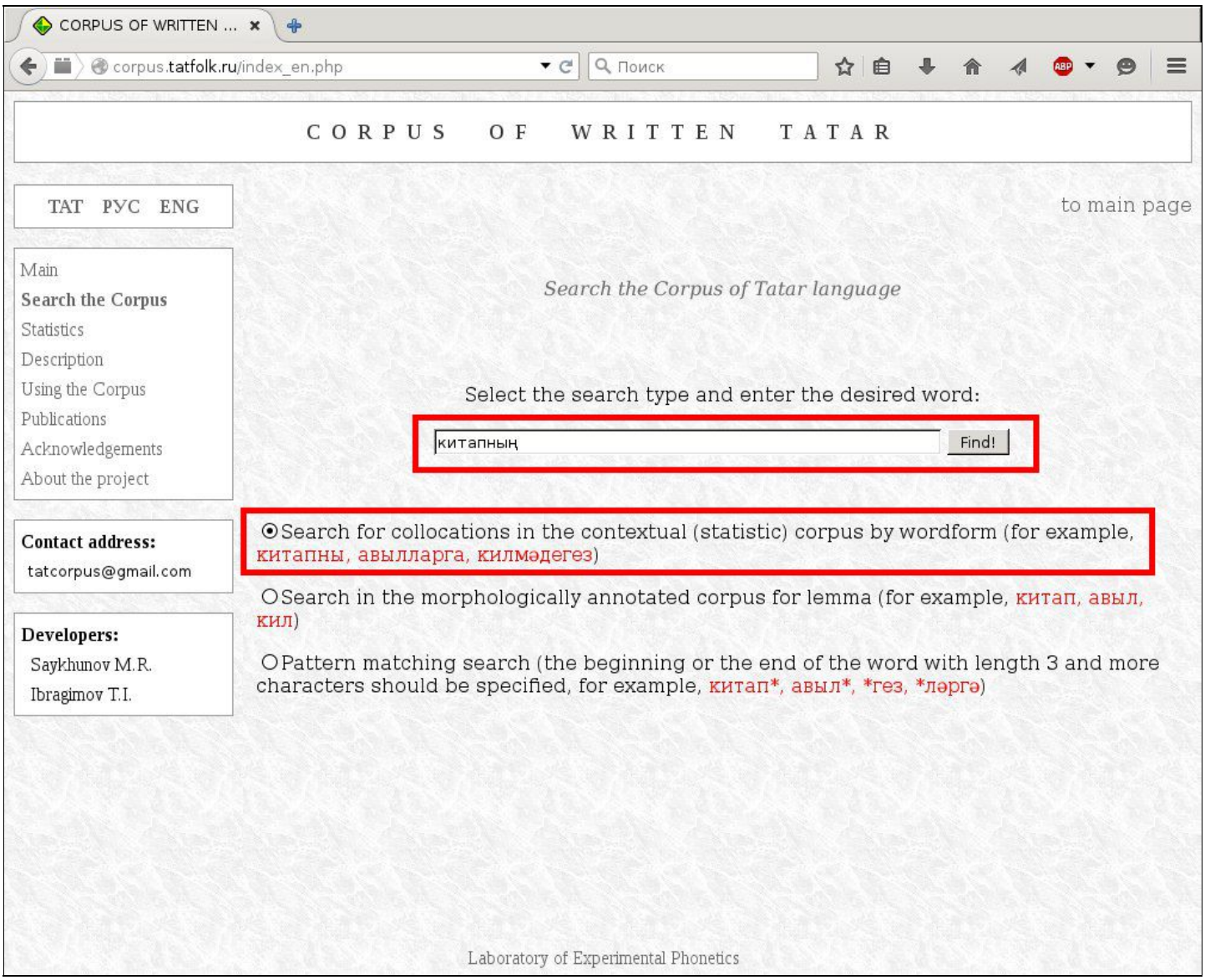


How to find a certain word combination in the Corpus?

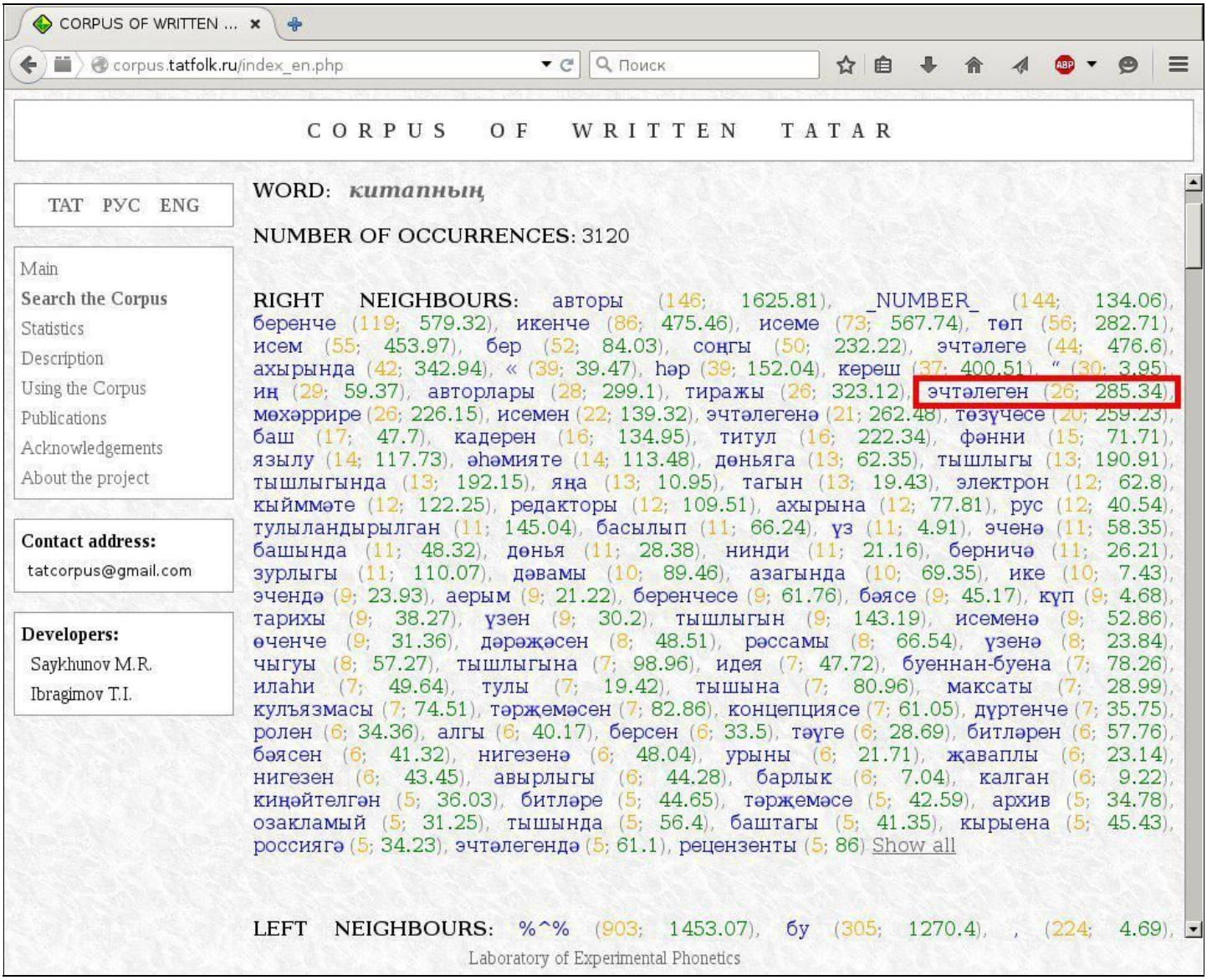
If you want to find occurrences of a certain word combination (collocation, phrase) in the texts, the contextual-statistical orientation of the Corpus will offer you ways to do the task.

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

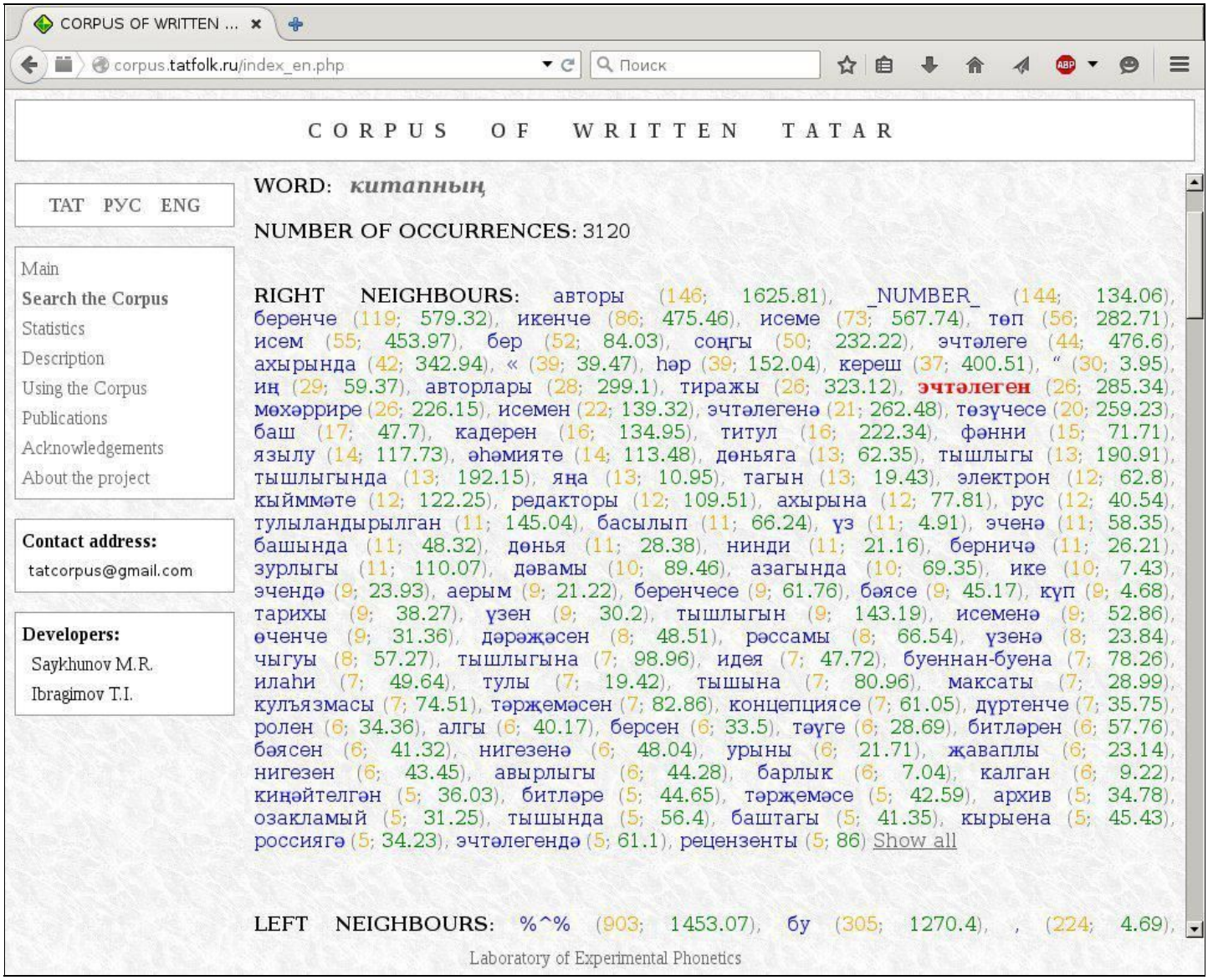
1) **Through the right neighbours.** First, type the word in the left position, i.e. «***китапның***», in the text field of the search page. Then select the checkbox named «**Search for collocations in the contextual (statistic) corpus by wordform...**»



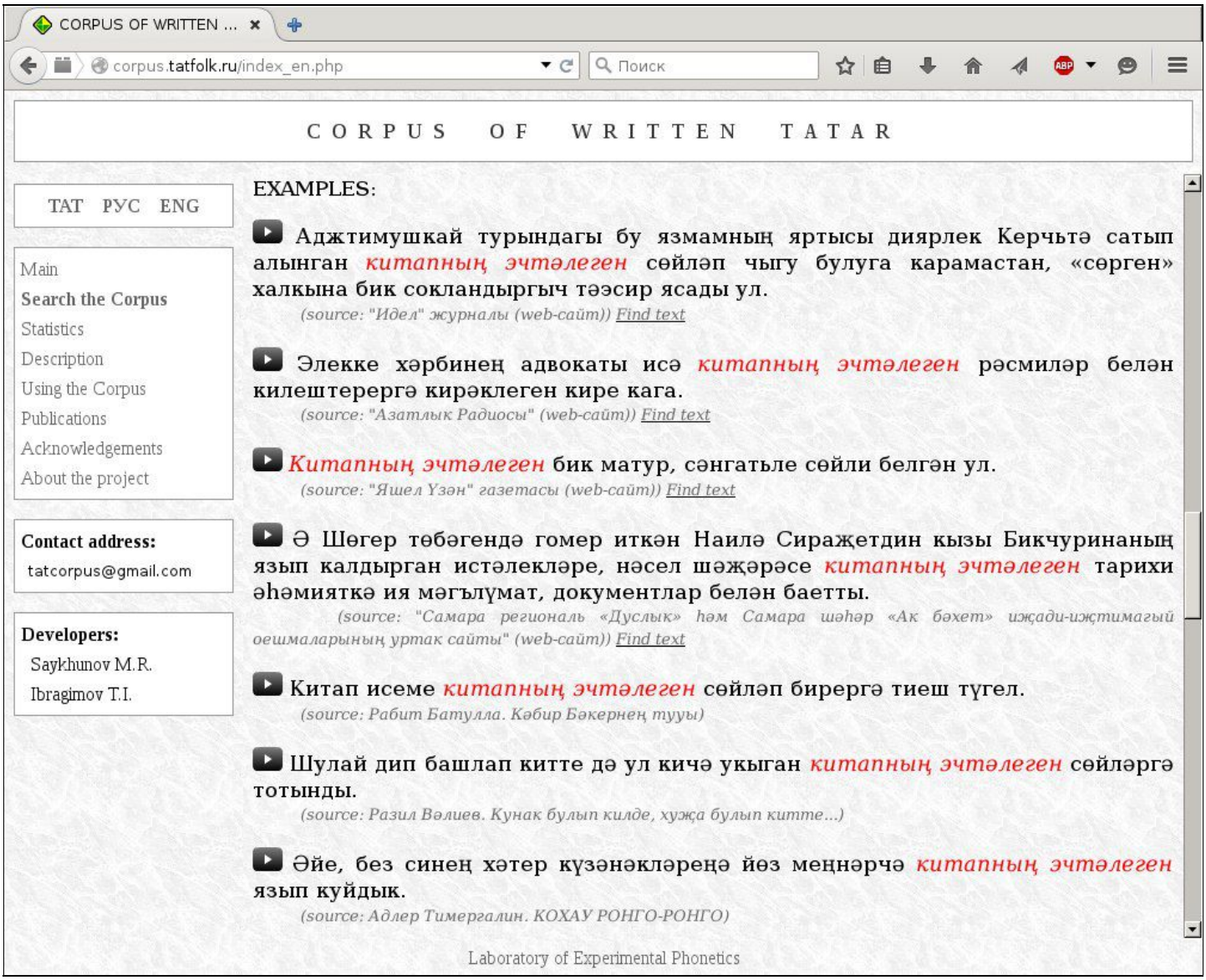
In the window now opened, find the word «***эчтәлеген***» among the "**Right neighbours**":



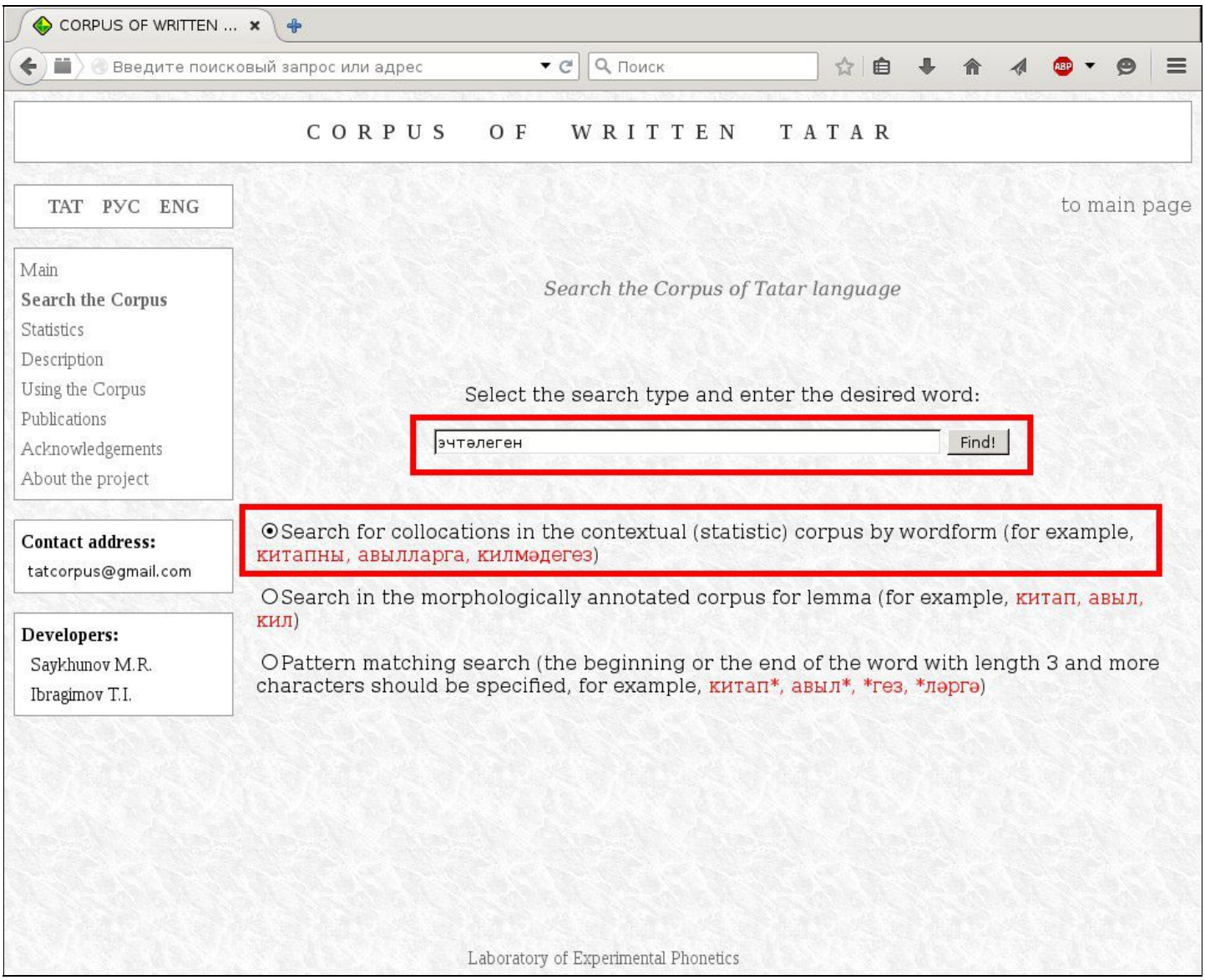
Click this word («***эчтәлеген***»), and see how it turns into highlighted red:



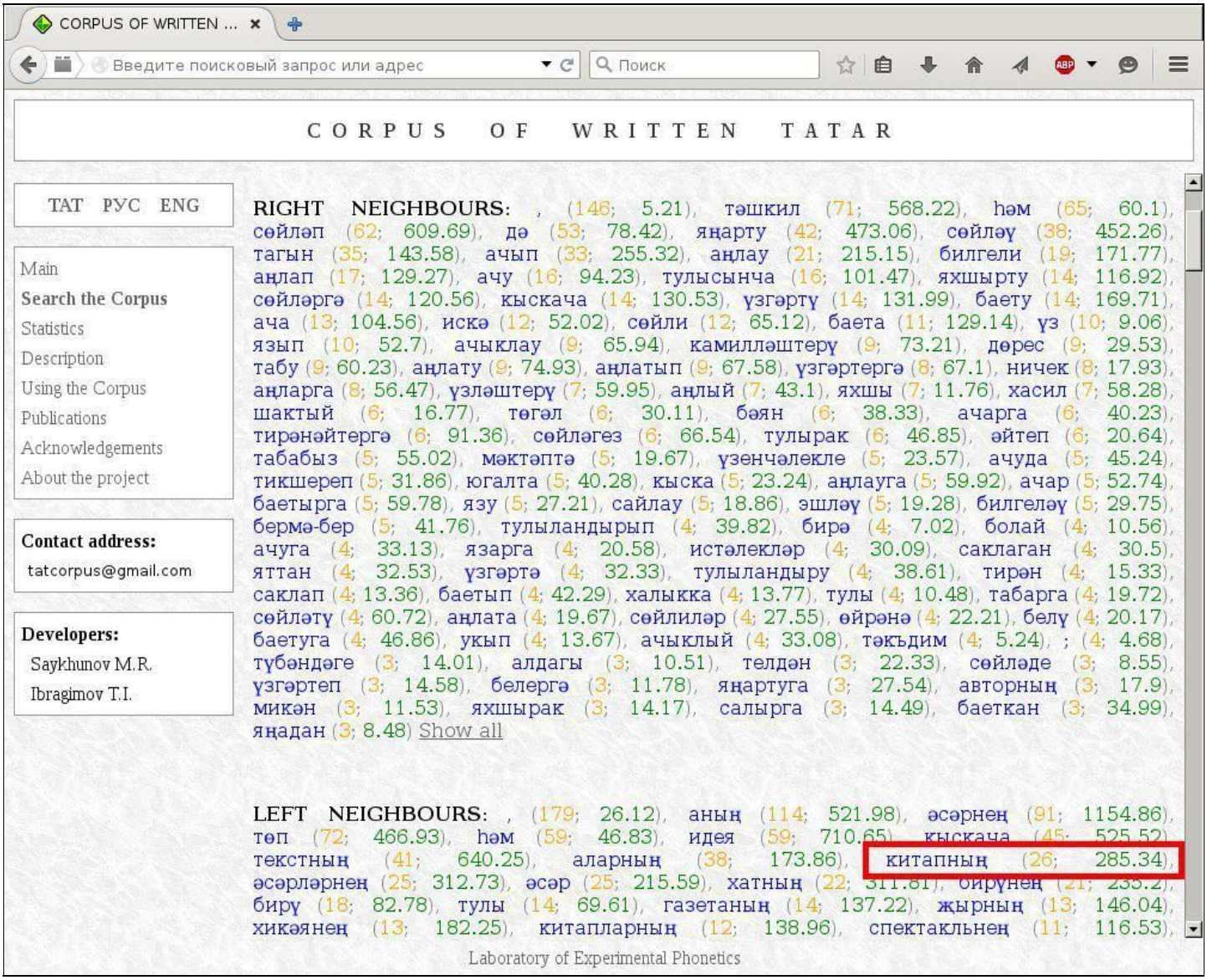
Below, in the section "**Examples**", there will be a list of sentences where this word combination is used:



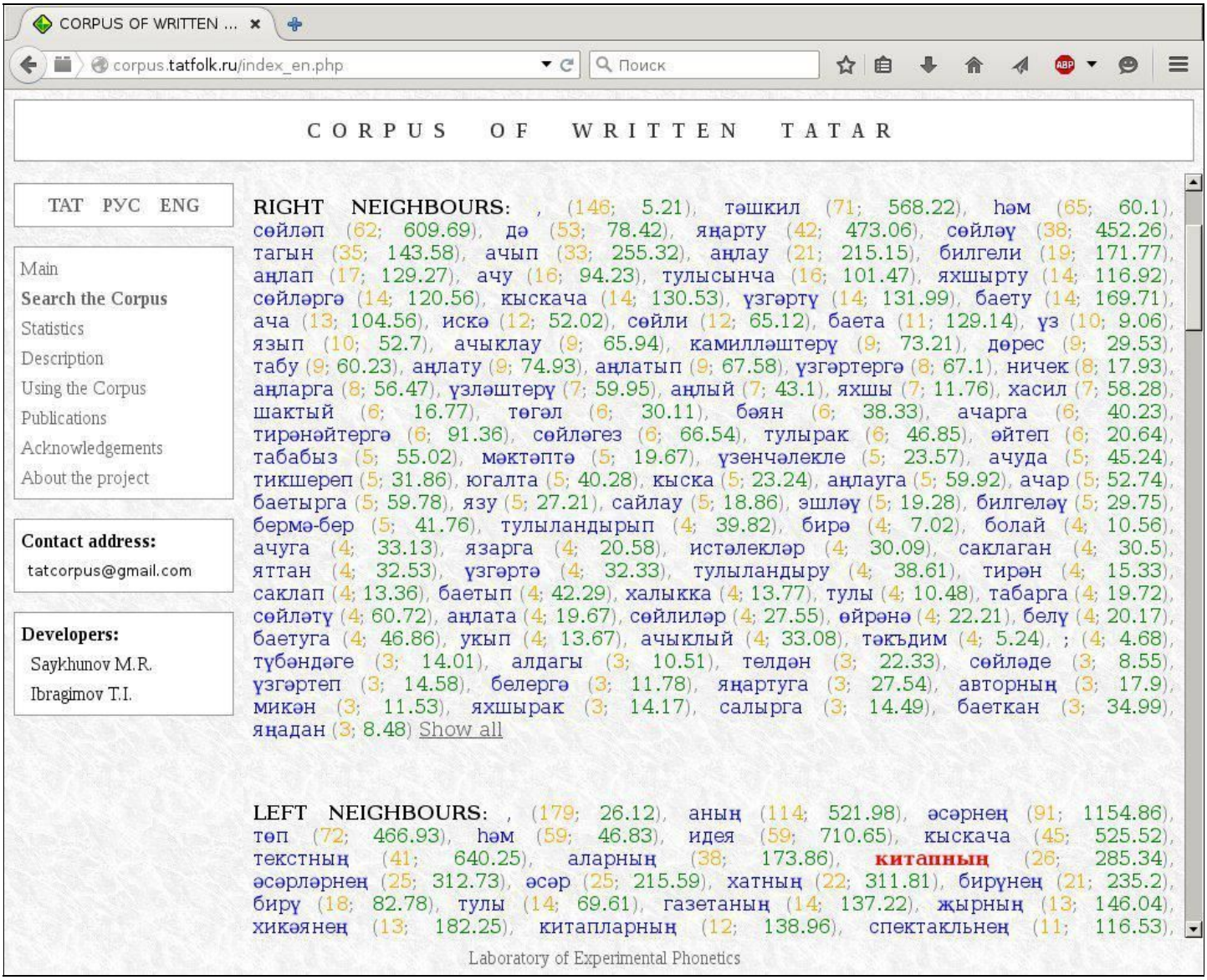
2) **Through the left neighbours.** If you use this alternative, write the right-side (latter) word in the text field on the search page, in the given example: «***эчтәлеген***».



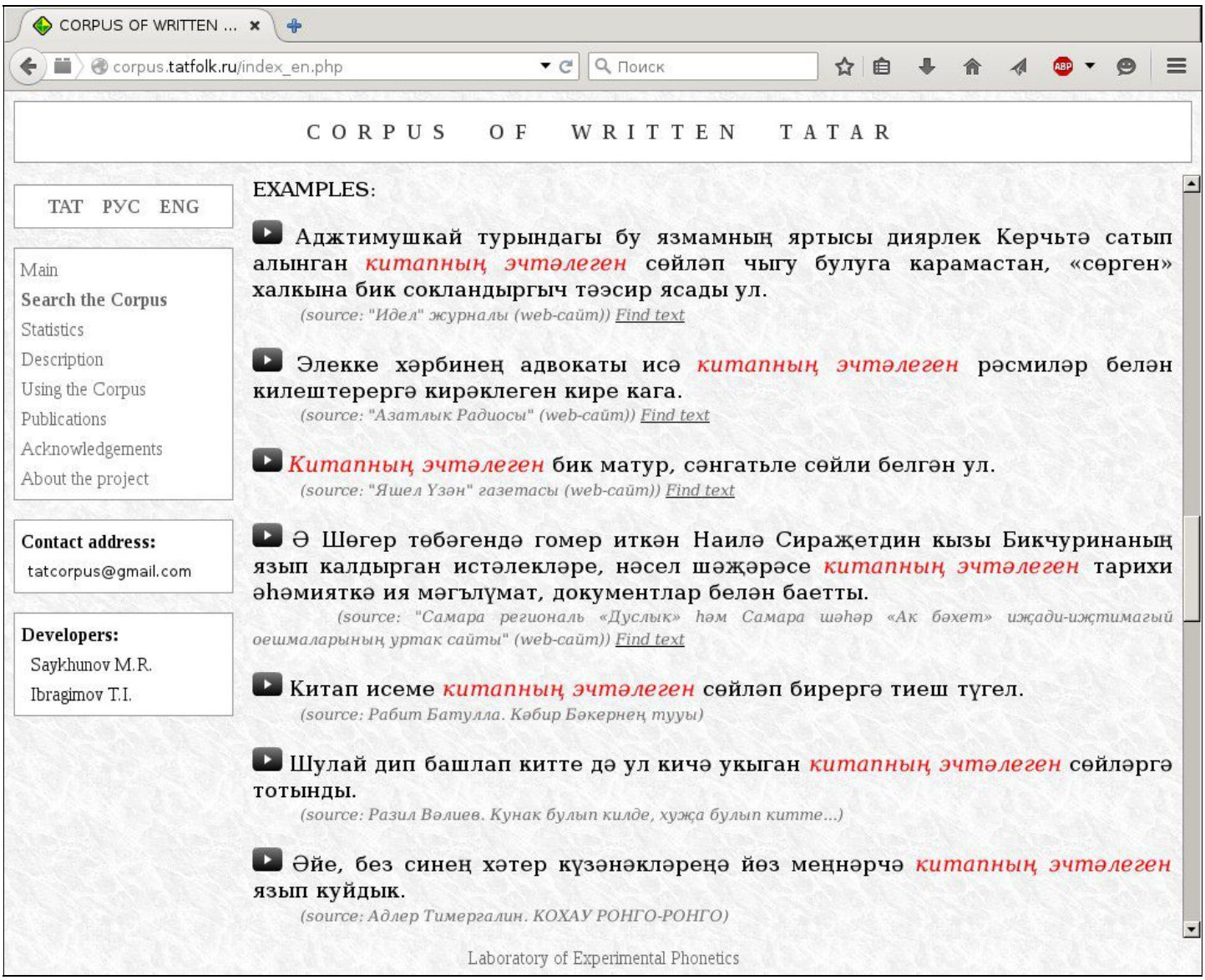
Among the "**Left neighbours**" you will find the word «***китапның***»:



After a mouse-click, the word turns red:



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.



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

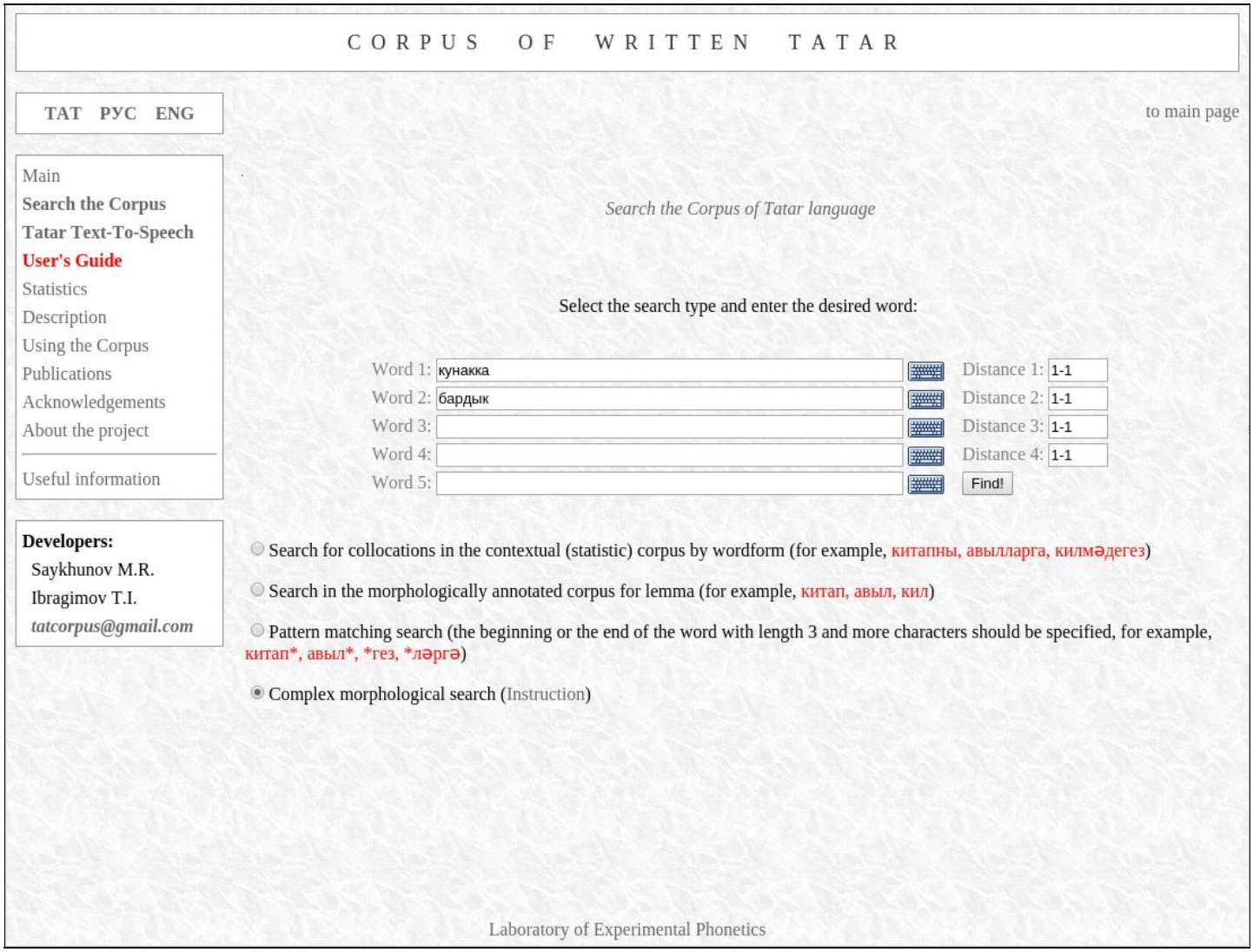
The morphological marking of the Corpus was carried out in 2014. The meta-language of grammatical labels is based on the system of tags for Turkic languages developed by the international project Apertium. This project aims to develop automatic translating system for a big variety of languages. The main 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:

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



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.

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.

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

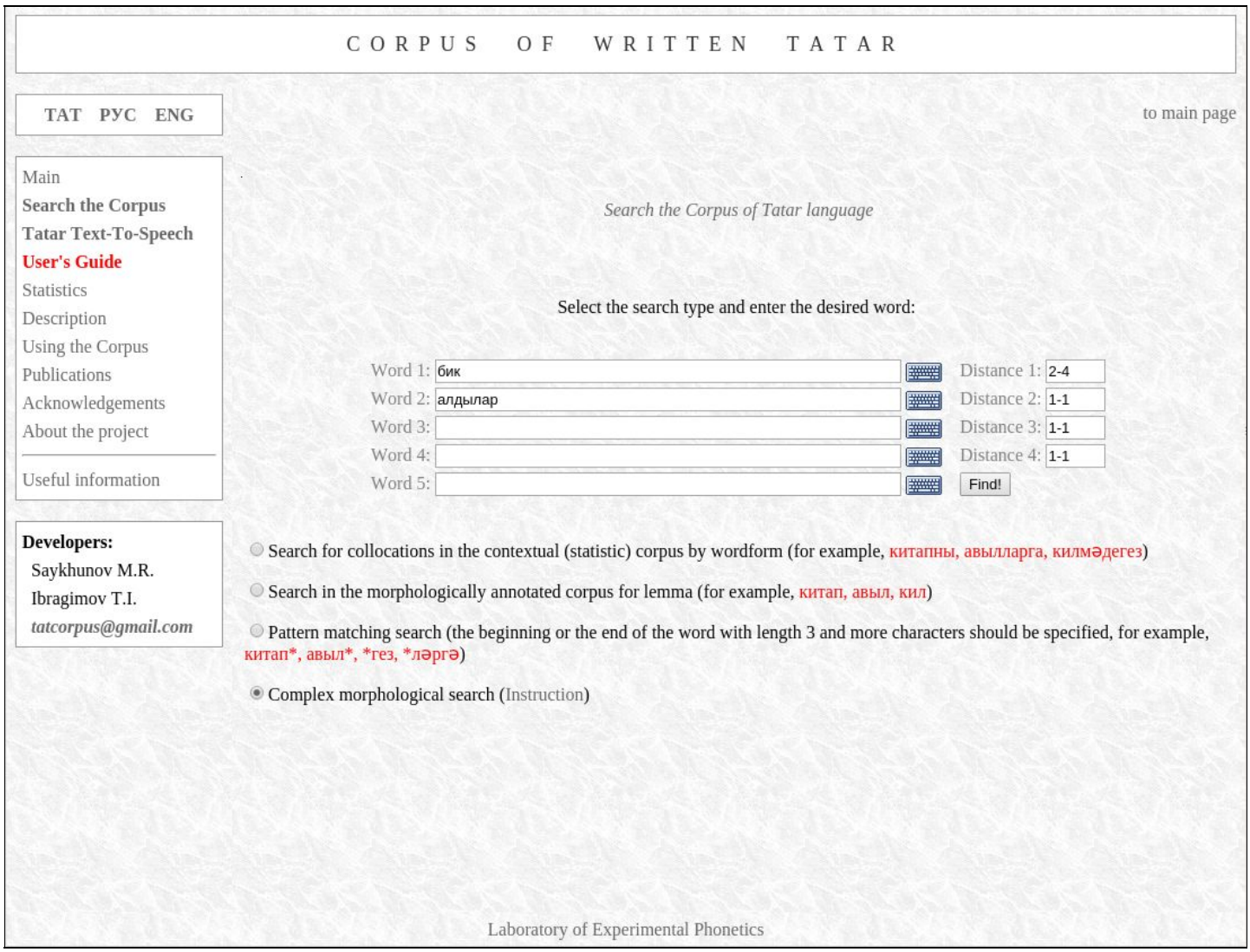
A note for the user: there might also be erroneous tags because the morphological annotation of the Corpus was made automatically.

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.

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

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.

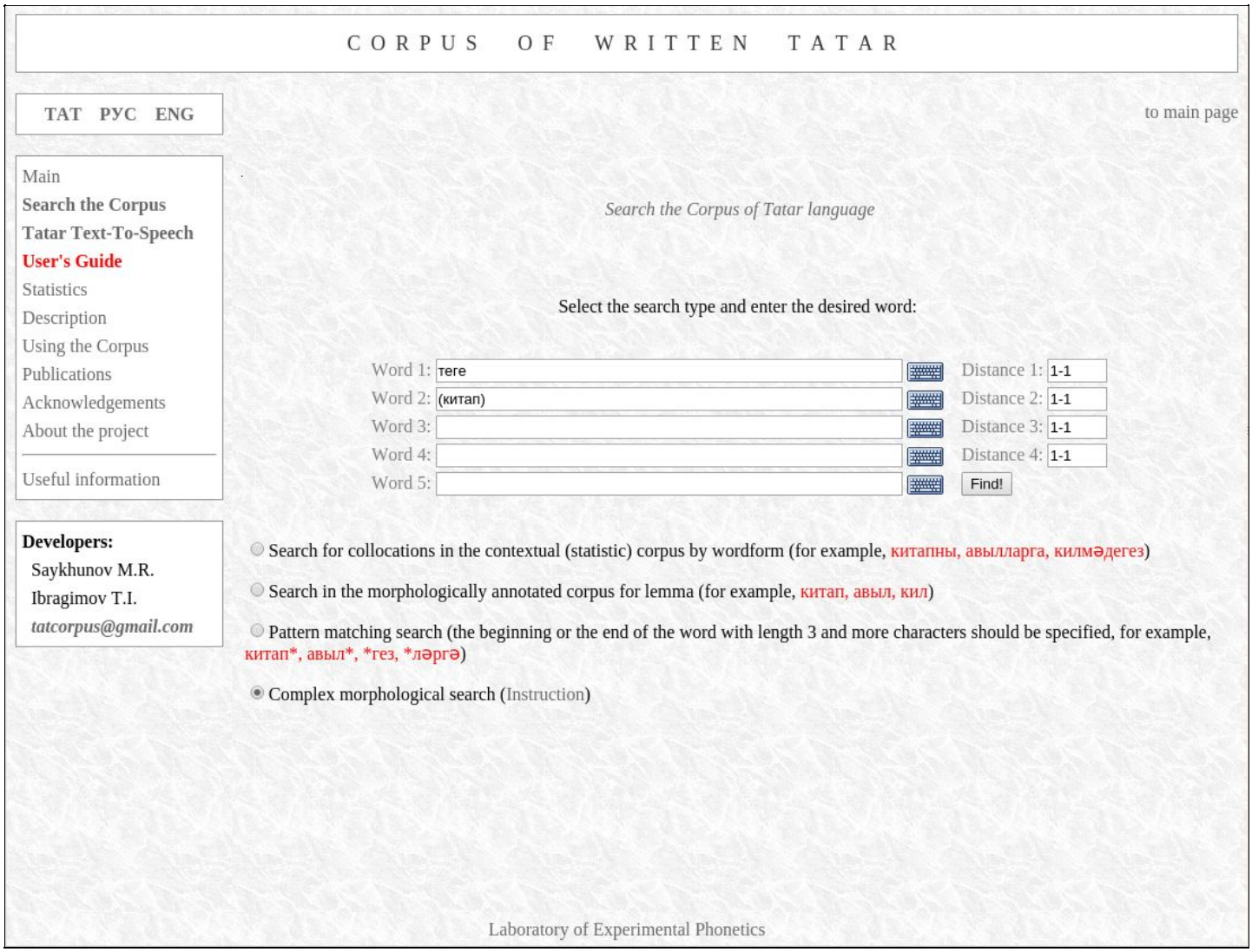


For example, the results of the search query «**{бик} 2-4 {алдылар}**» begin as follows:

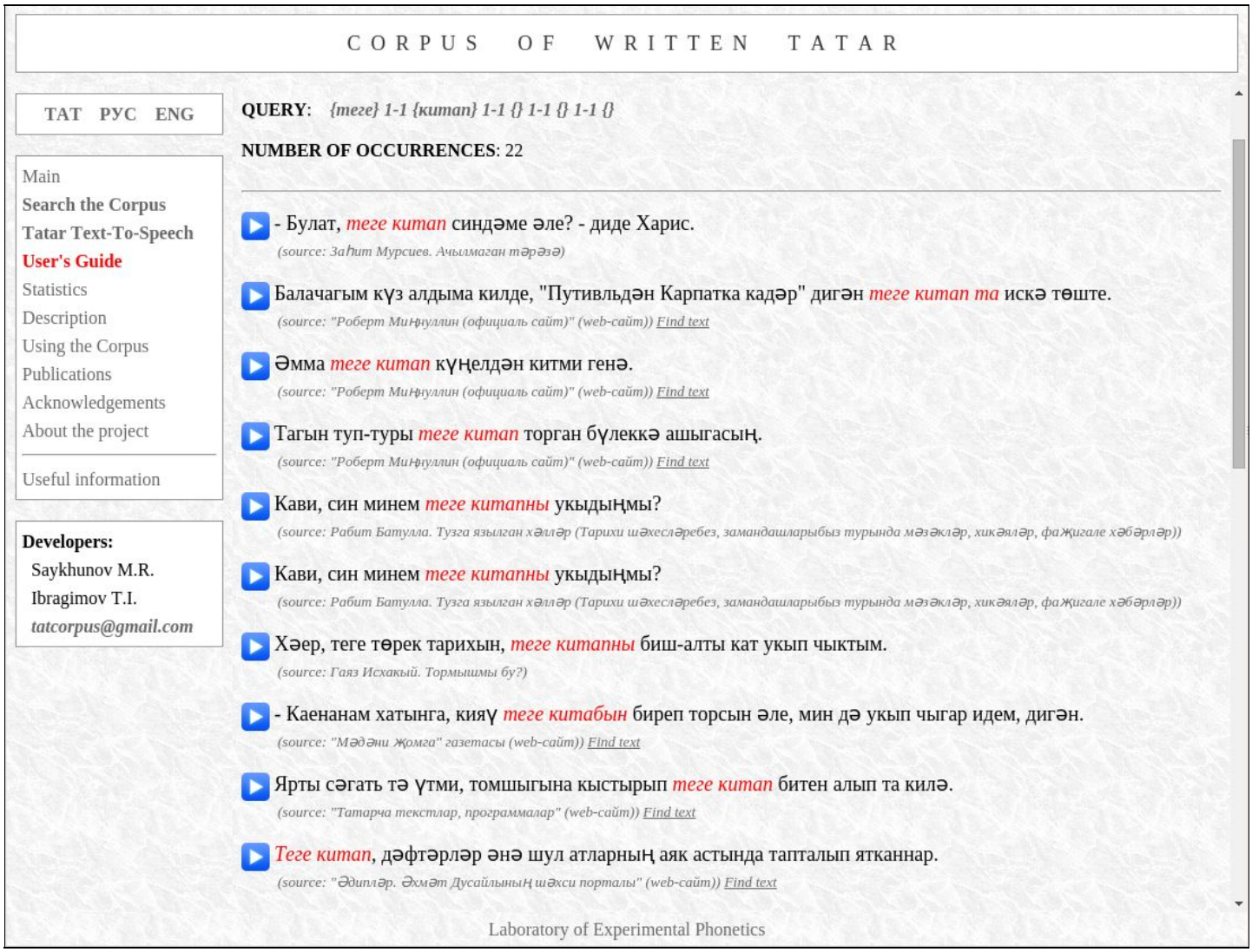
How to find combinations using lemmas?

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, «***(китап)***». 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.

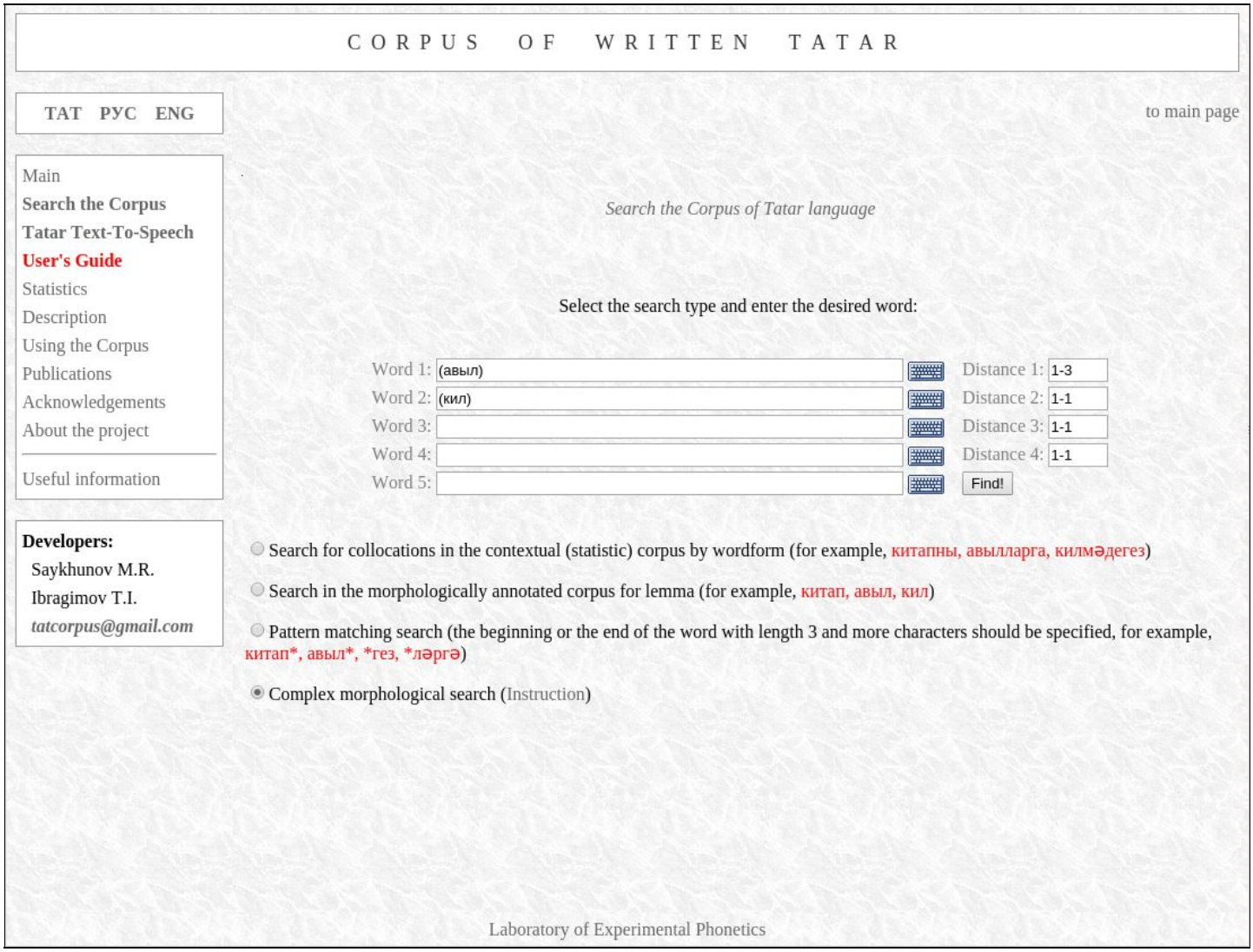
Let us perform the query «**{теге} {(китап)}**» (“*that (book)*”):



The results can be seen below:



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.



The results are as follows:

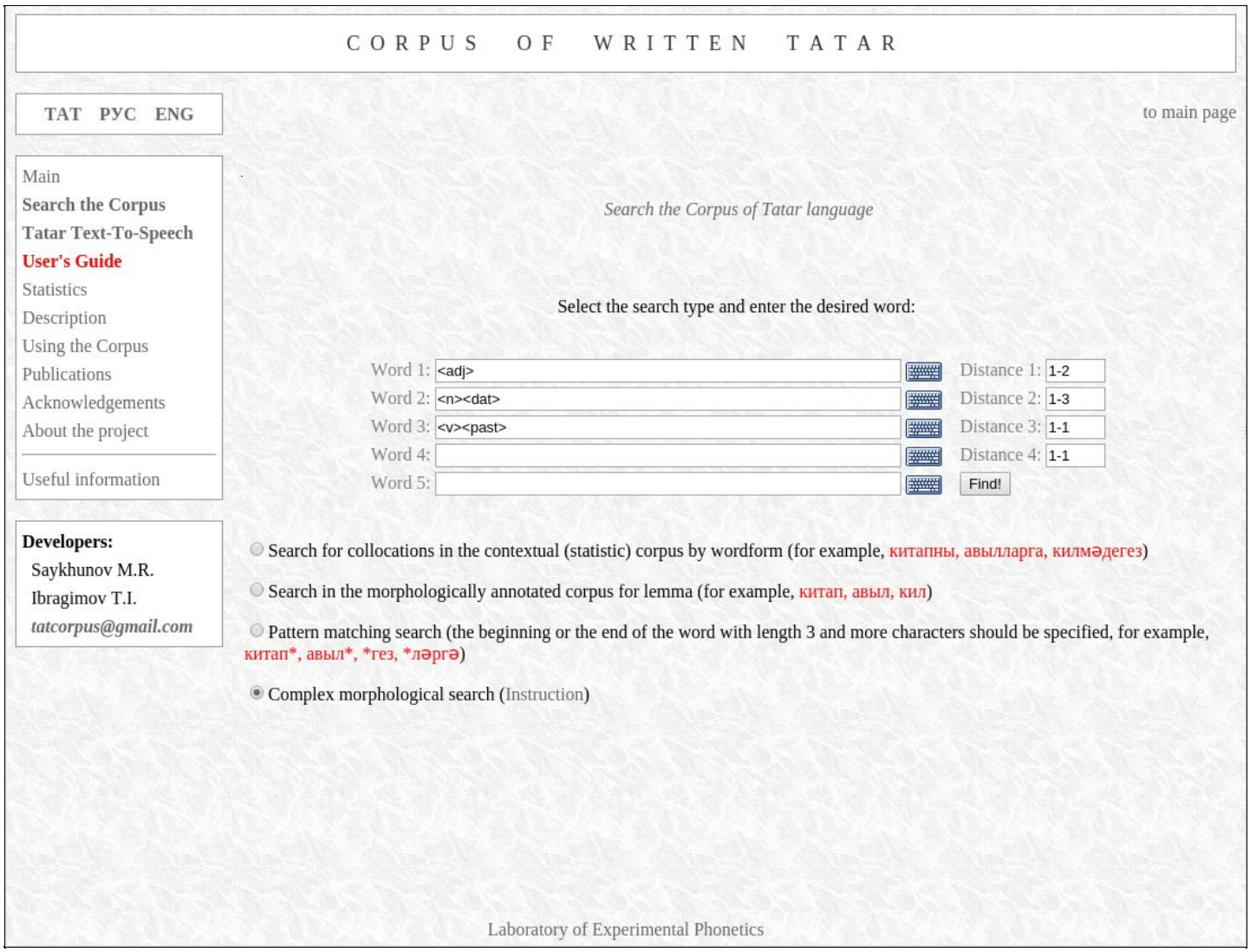
How to find combinations using morphological (grammatical) tags?

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

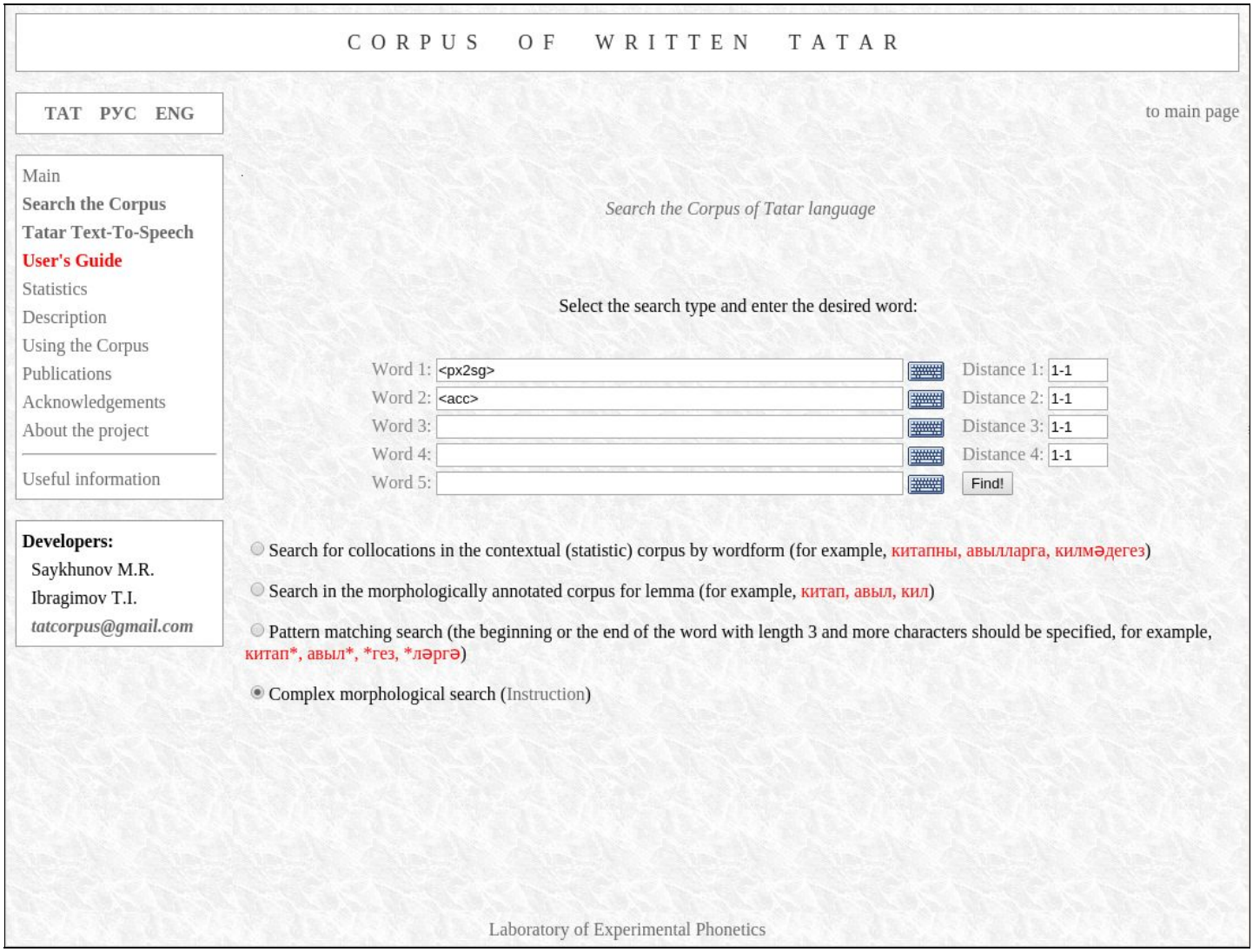
In the search form, one can define one tag or several tags. Only the combination type «and» is allowed, not «or». This means that it is not possible to use definitions consisting of incompatible (exceptive) tags like in «**<dat><acc>**» (*dative case, accusative case*) in a search targeting occurrences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



The results of the search begin with the following:

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>**).



The results can be seen below:

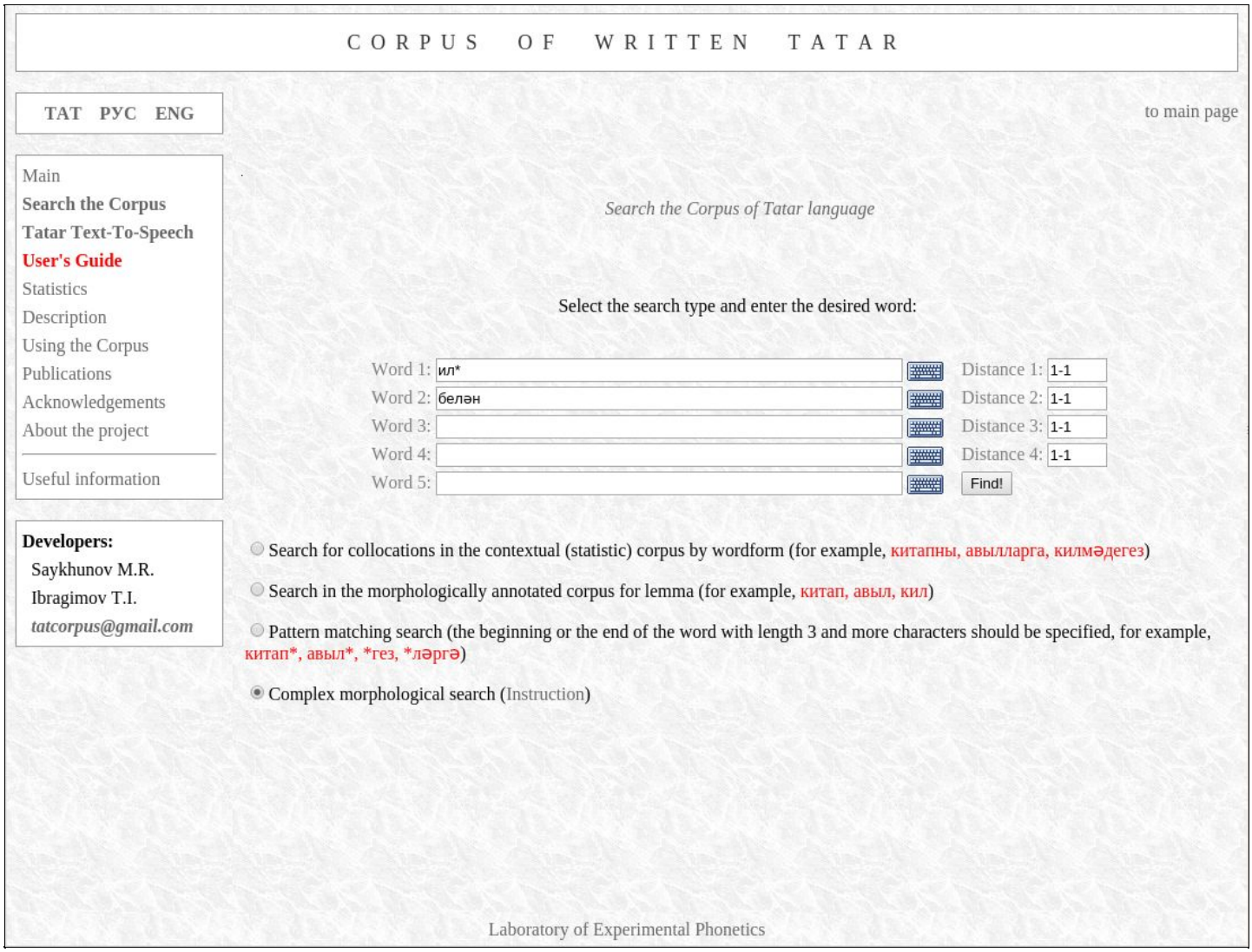
How to find combinations using the beginning, the middle and/or the end of words?

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...**».

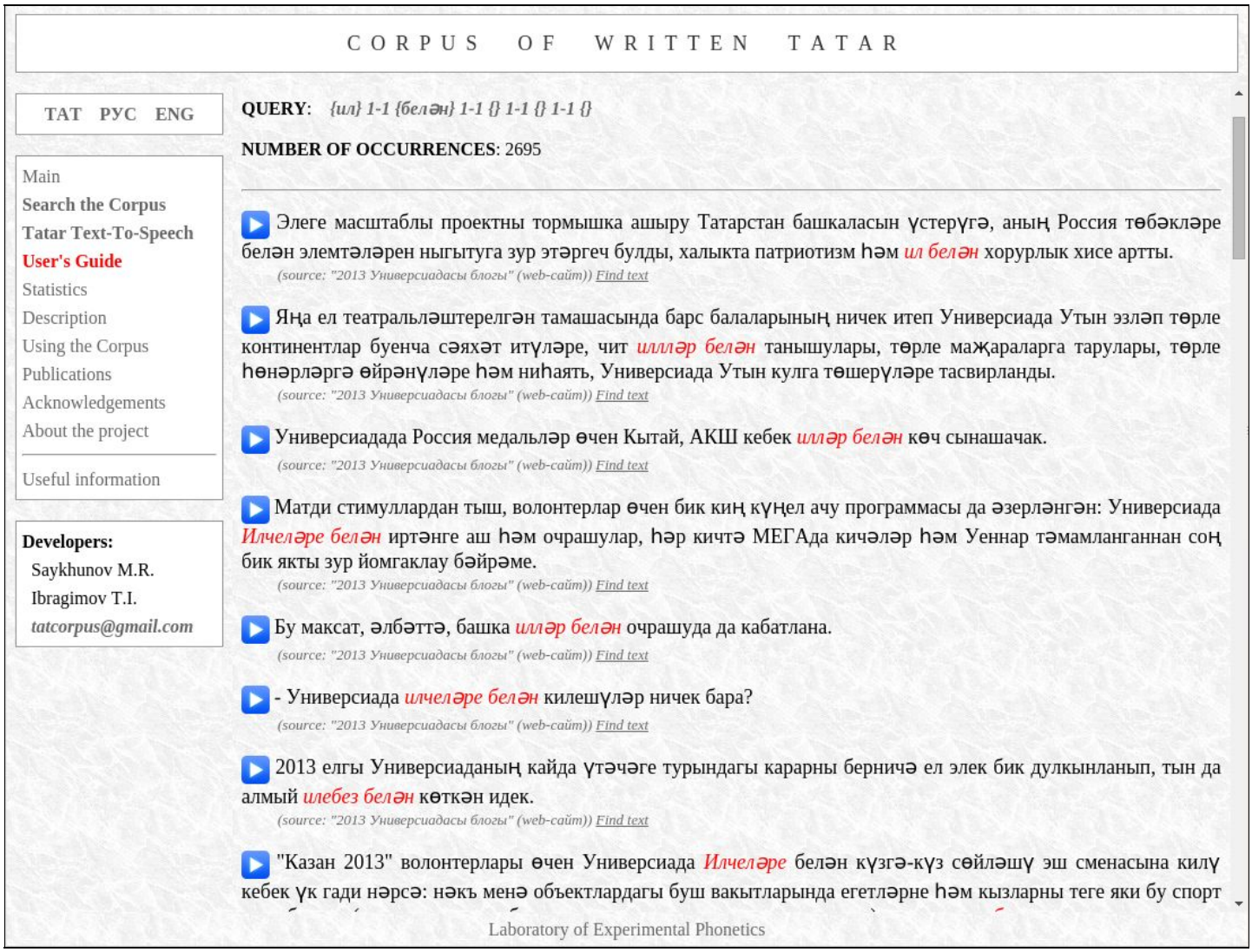
You can target the middle part of the word by using a query like «***әме***», which produces **керәмен, әмер, үсәме...**

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*).



We get the following results:

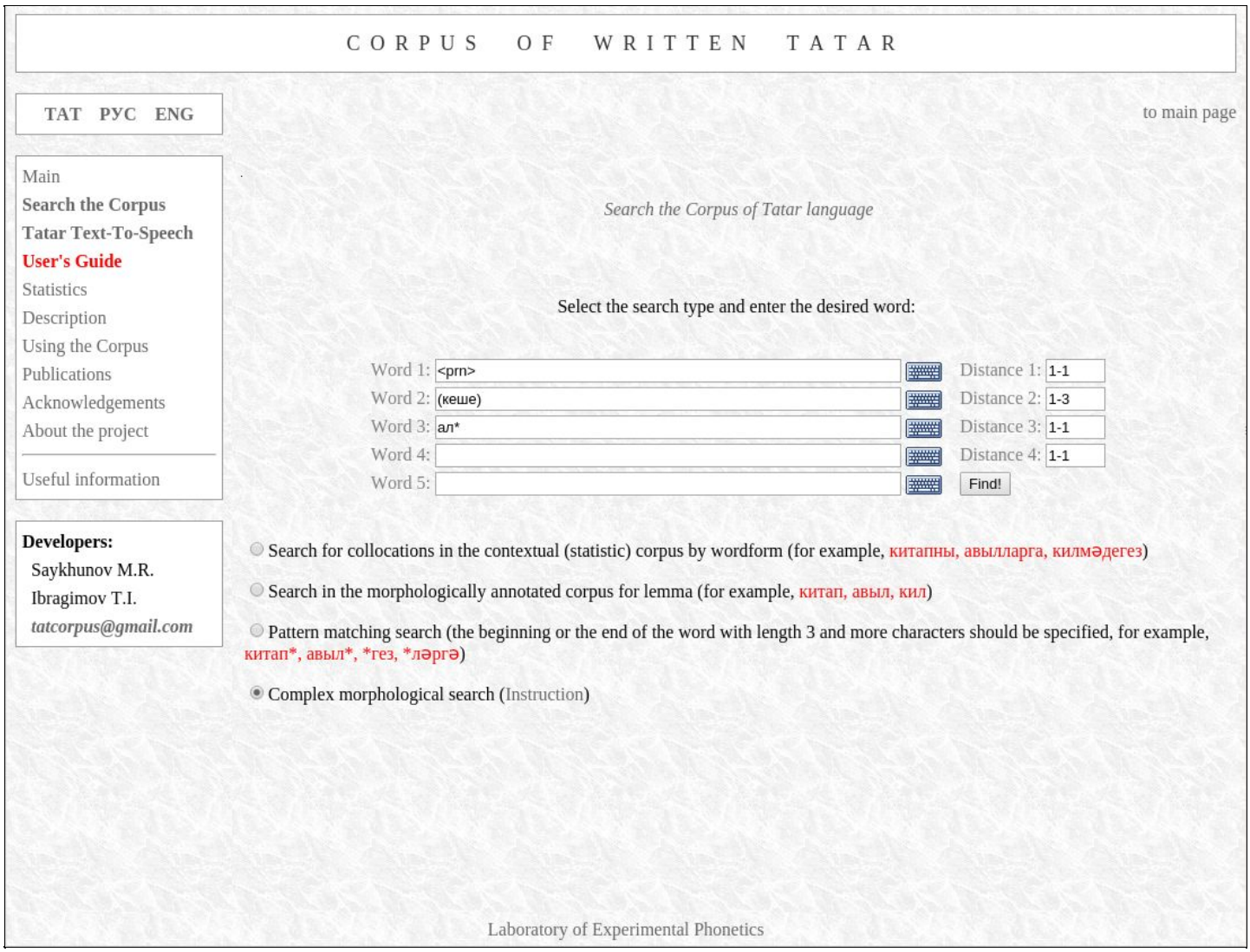


To make a search by the beginning and end of the word, you should place * mark between them, e.g. «**ал*лар**» produces words like «**алалар, алдылар, алмалар, алмагачлар...**».

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**».

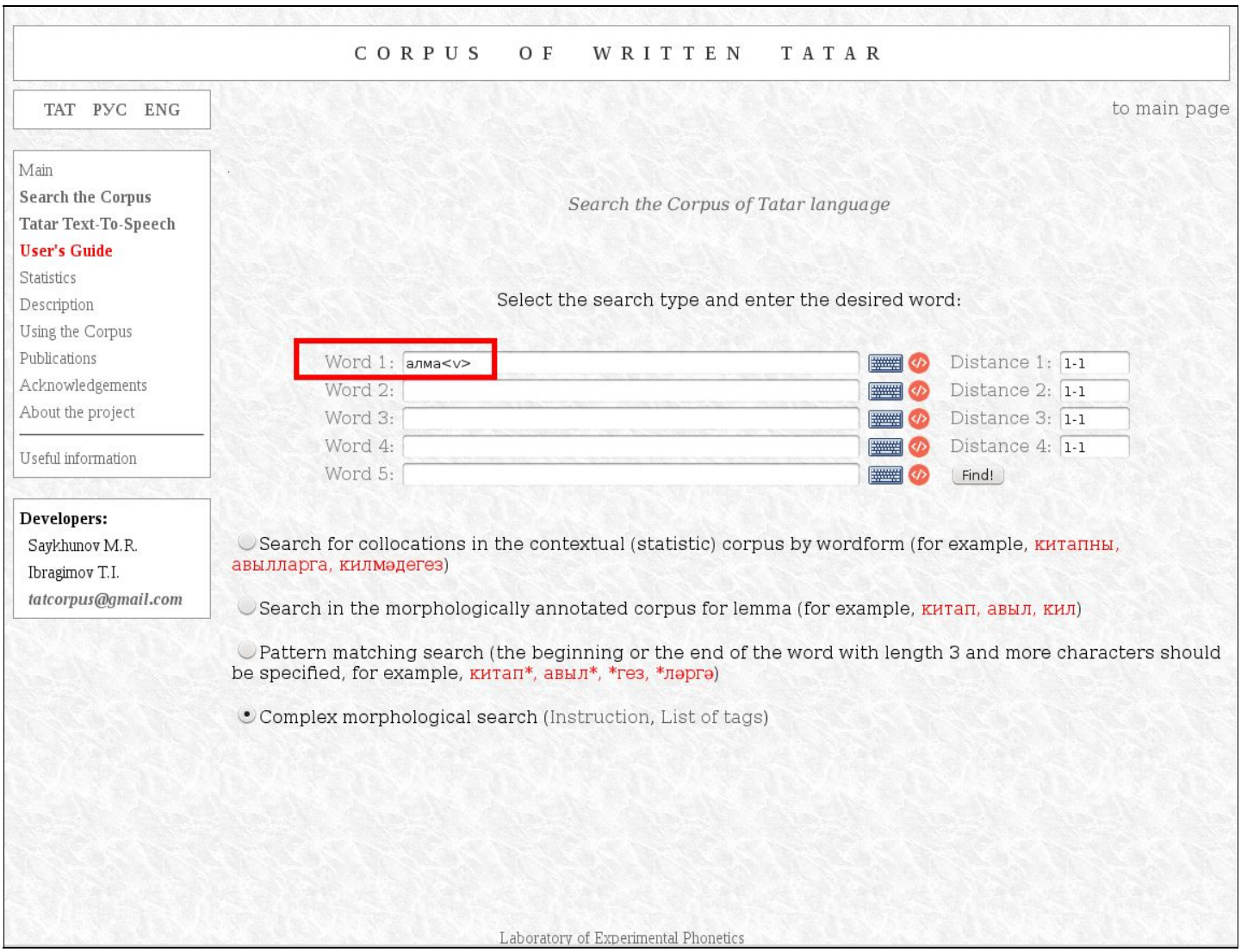


The query produces the following results:

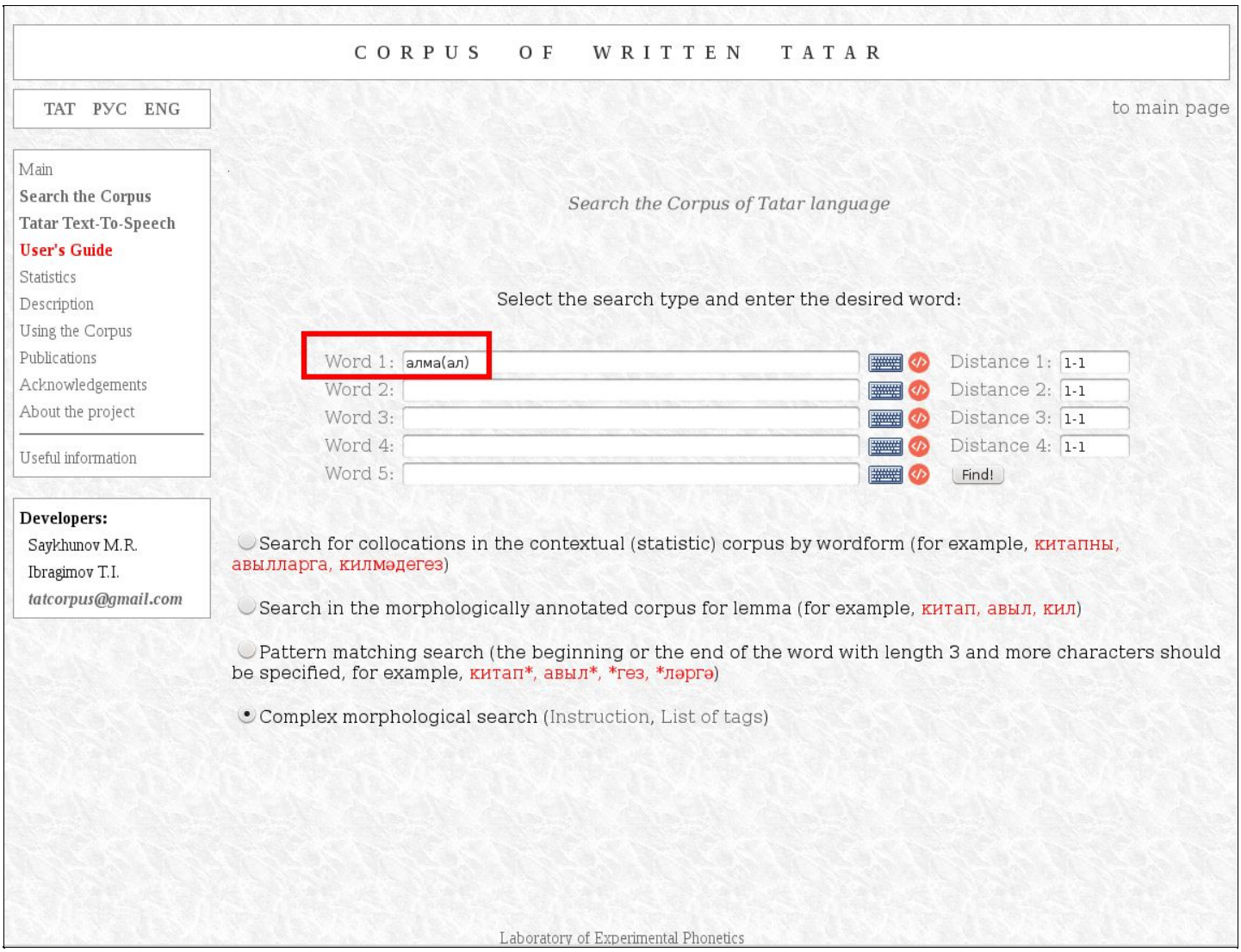
How to find combinations using several parameters for every word?

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>**».

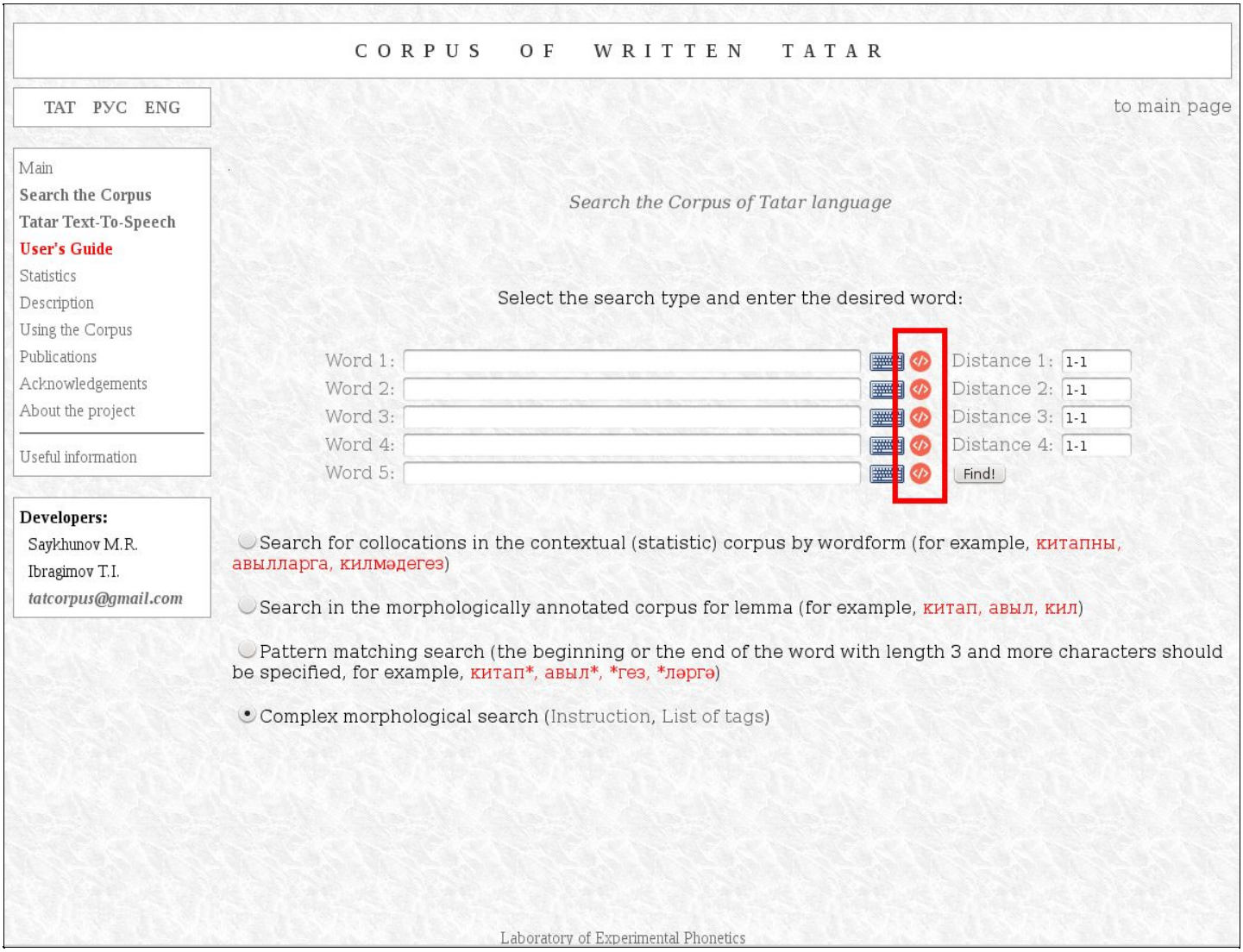


In this case, there is another way to solve the problem. You can place the right lemma, i.e. «**(ал)**» (*to take*), immediately after (or before) the word form «**алма**», resulting in the search definition «**алма(ал)**». So the system will search for those cases of «**алма**» where the lemma of the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

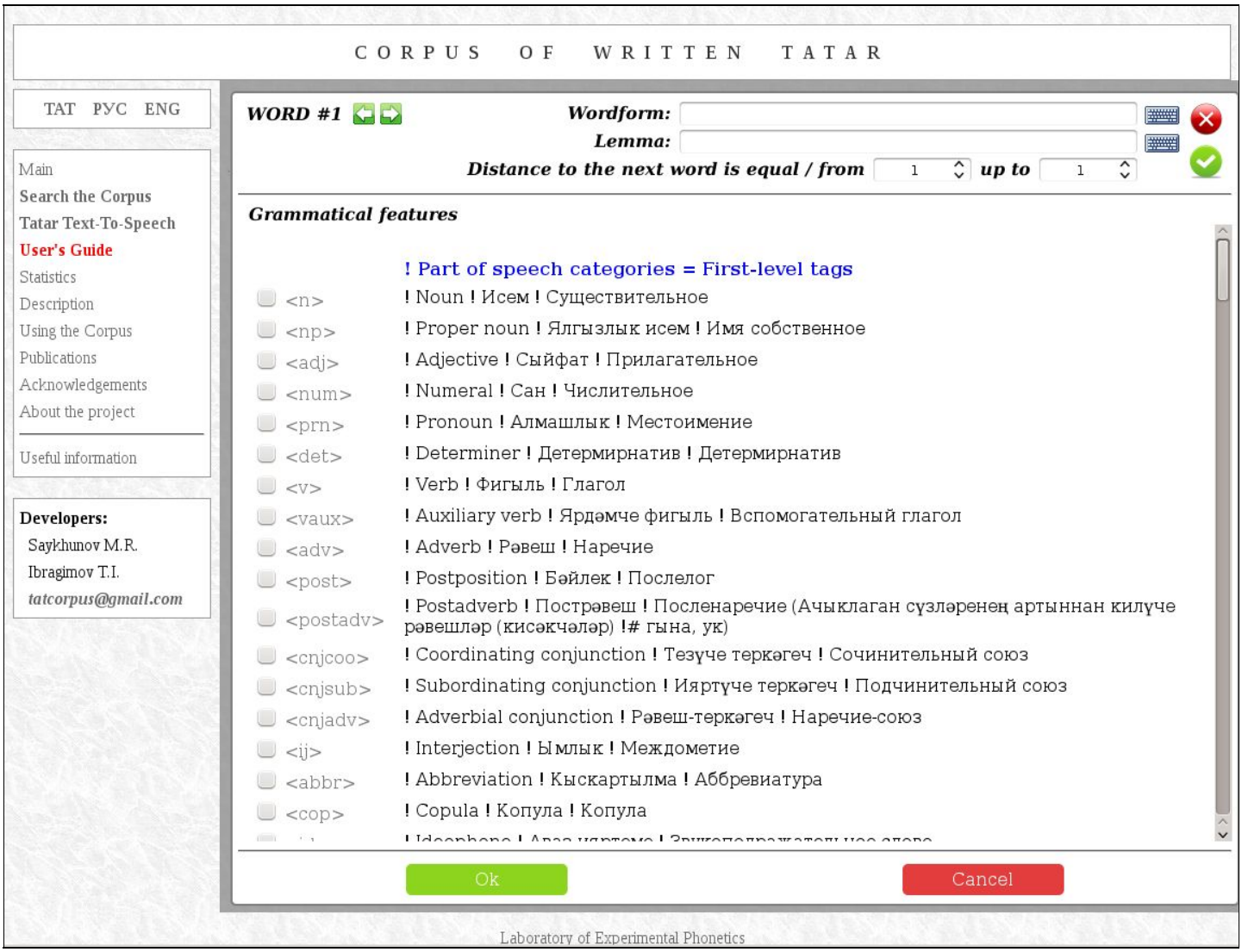


Entering search parameters in the graphical mode

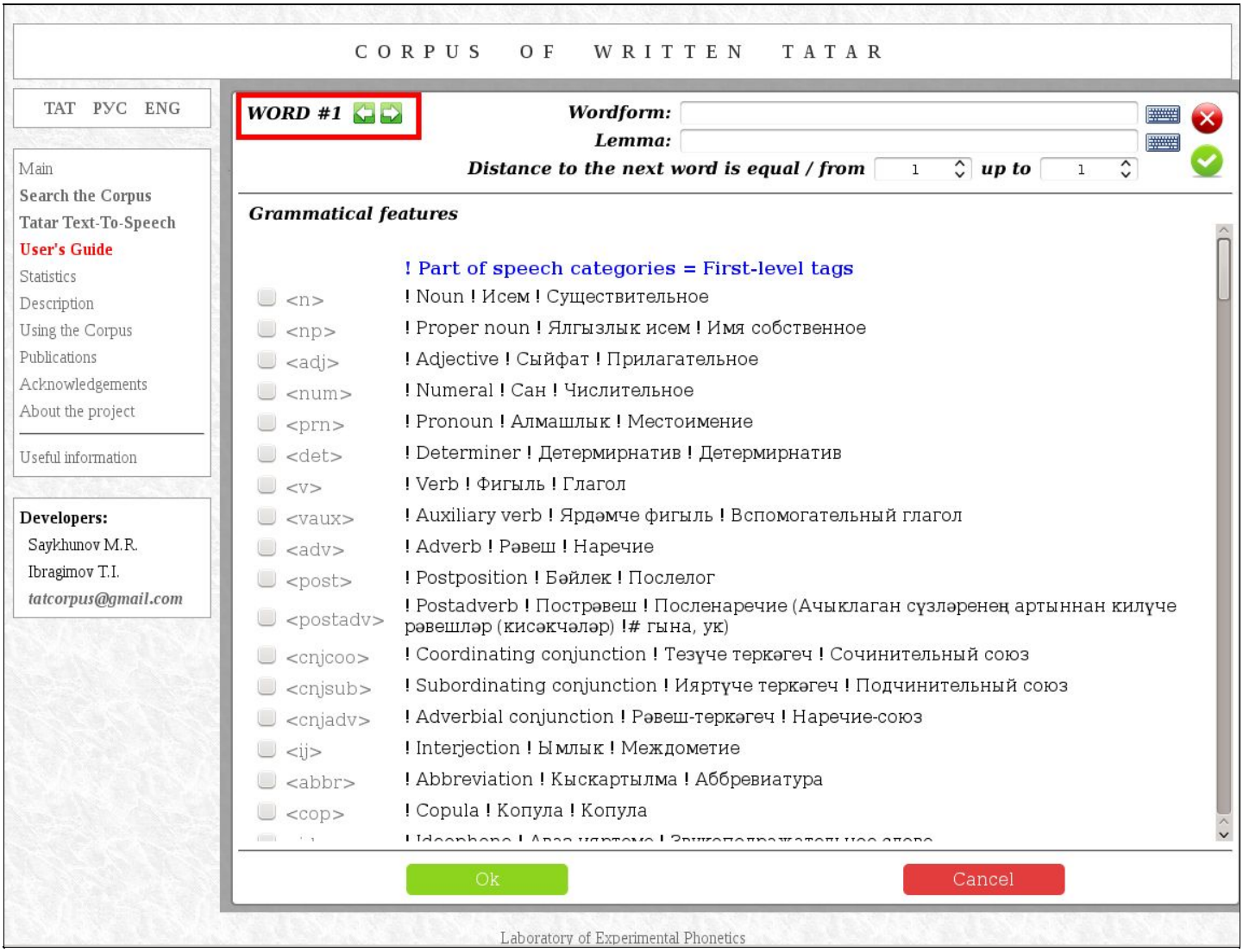
A special tool facilitates entering of the search parameters. You can access the tool by clicking the buttons to the right of the text fields, as shown below.



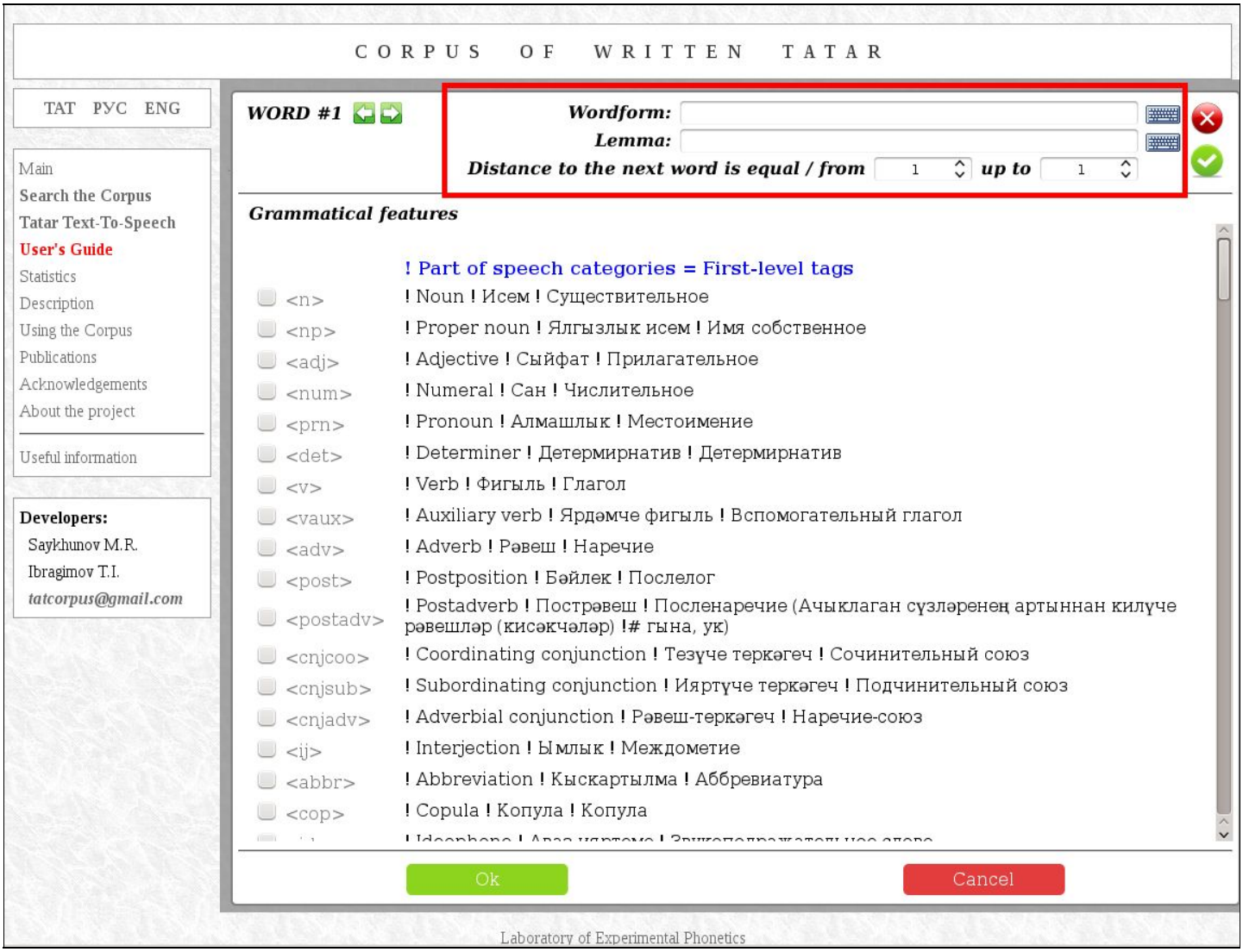
A window of the following shape will pop up.



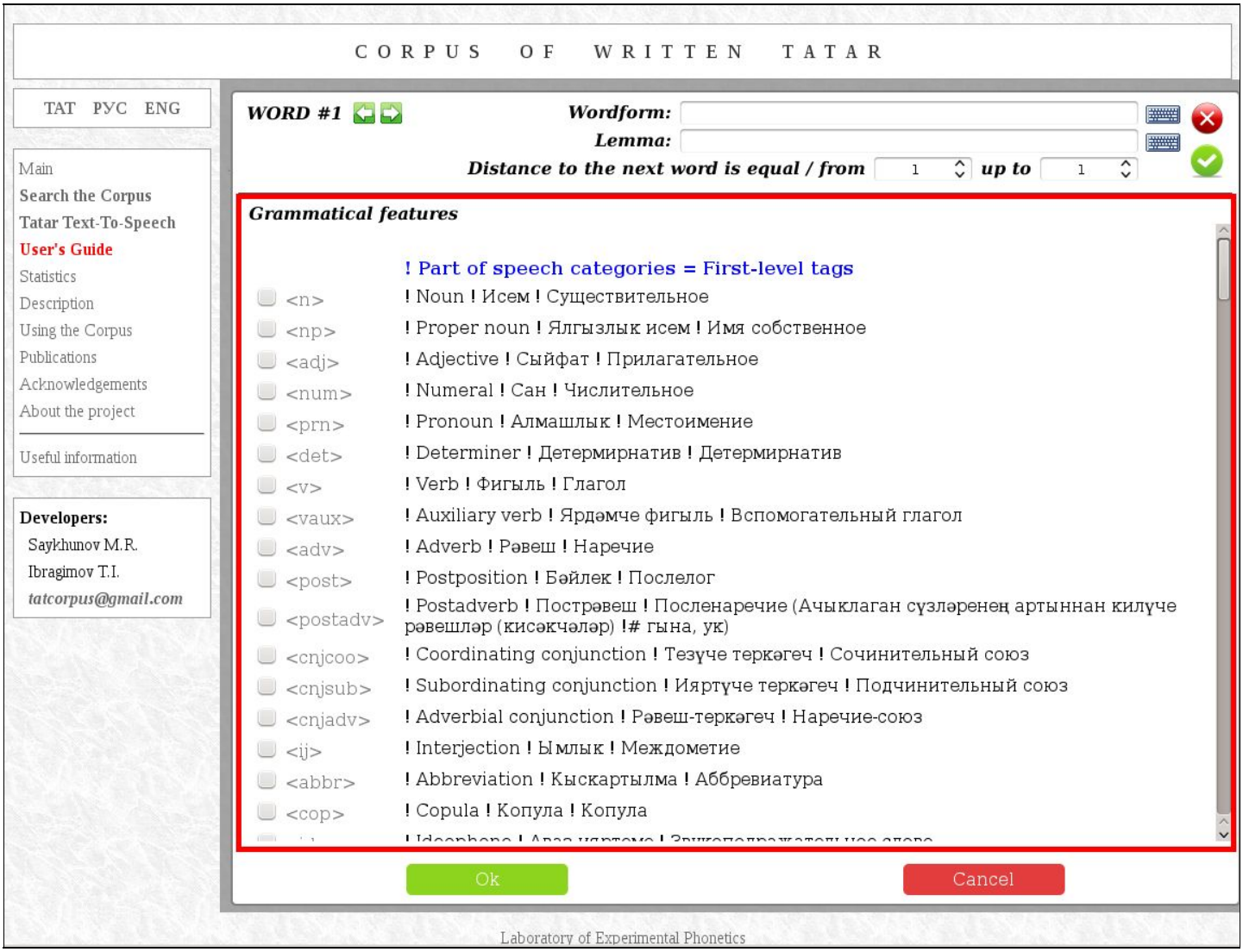
In the upper left corner, you can see the number of the current word. On the right side of the word, there are two buttons that allow you to quickly jump to editing the previous or the next word. The information entered into the current page will be automatically saved, when you switch between pages!



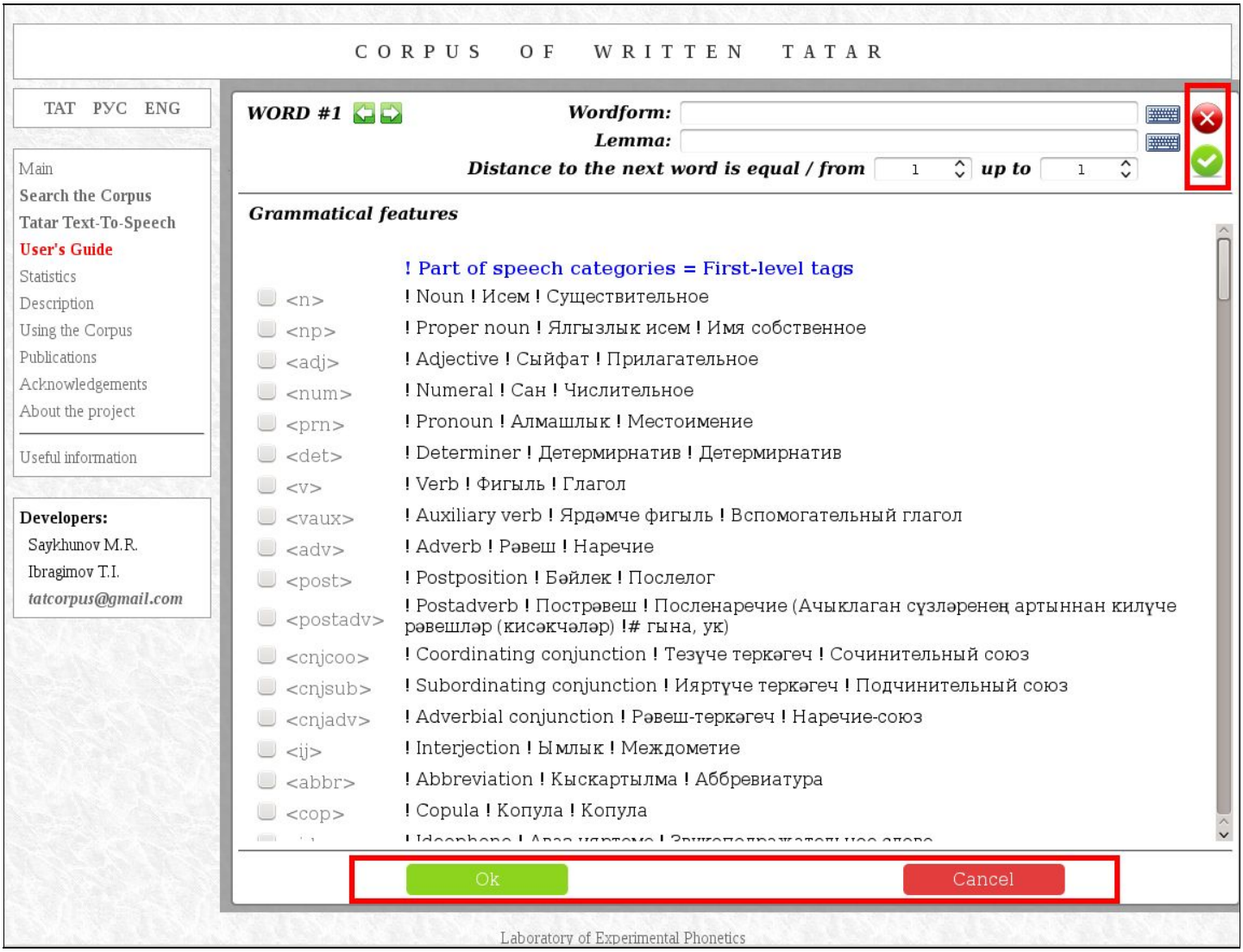
The red highlighted area on the picture that shows where you can specify the word form, lemma and interval. Please note that here the lemma must be entered without parentheses!



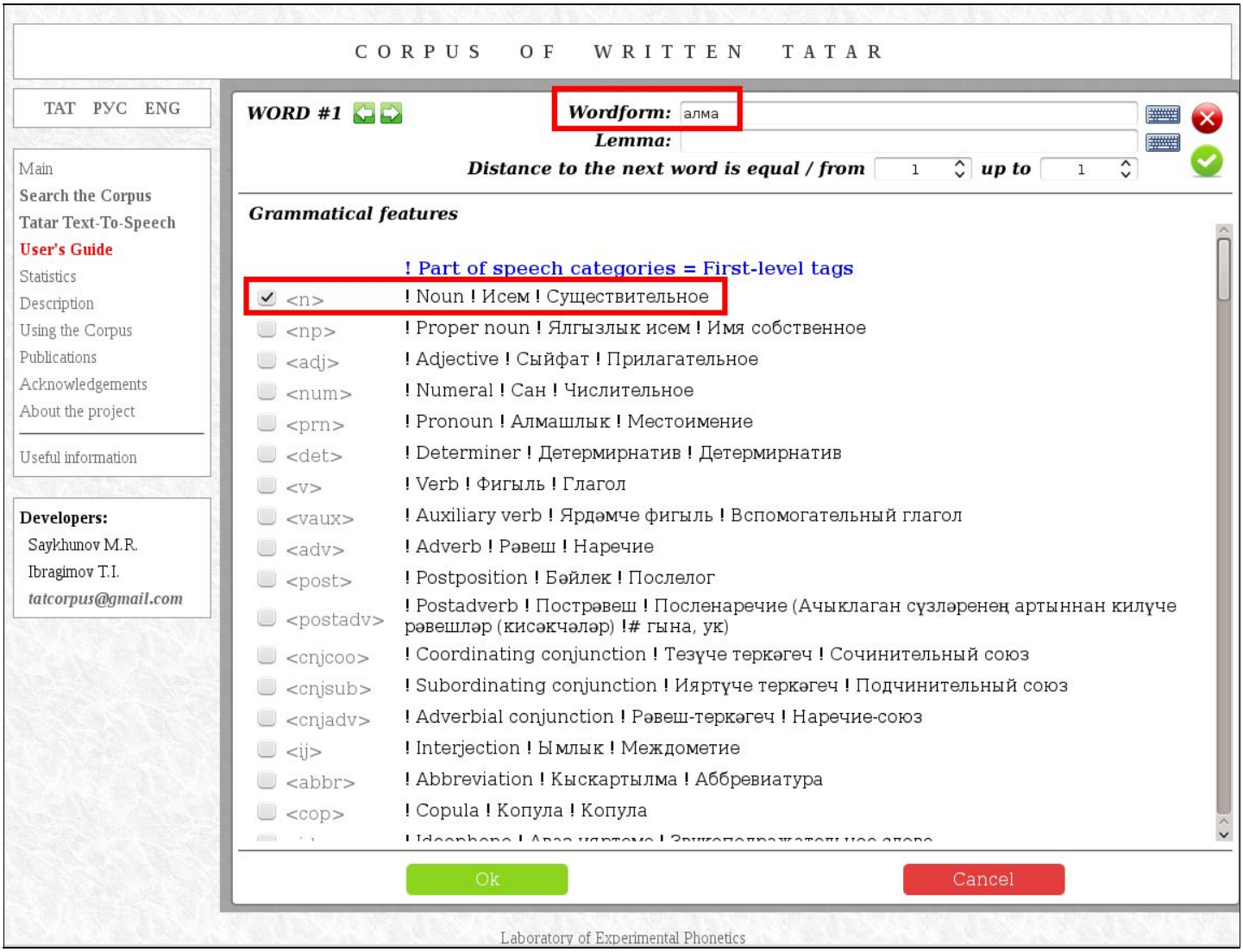
The main area of the window (highlighted in red) is where you can choose grammatical features to be included in the search pattern. The tags are listed by category. You can choose a particular tag by checking the box in the corresponding line.



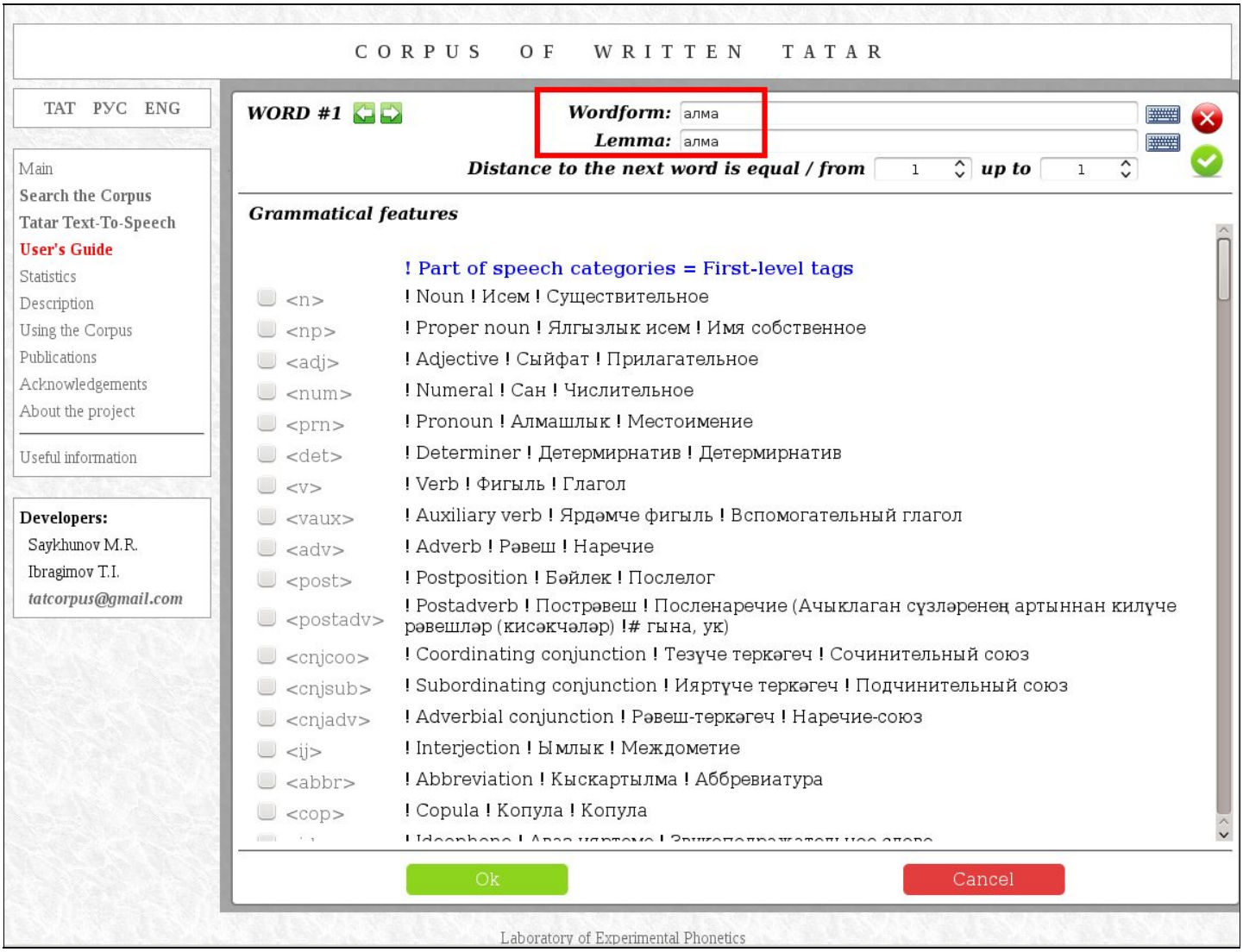
The buttons in the upper right corner and in the bottom of the page allow you to close the window with saving (green) or without saving (red) the changes.



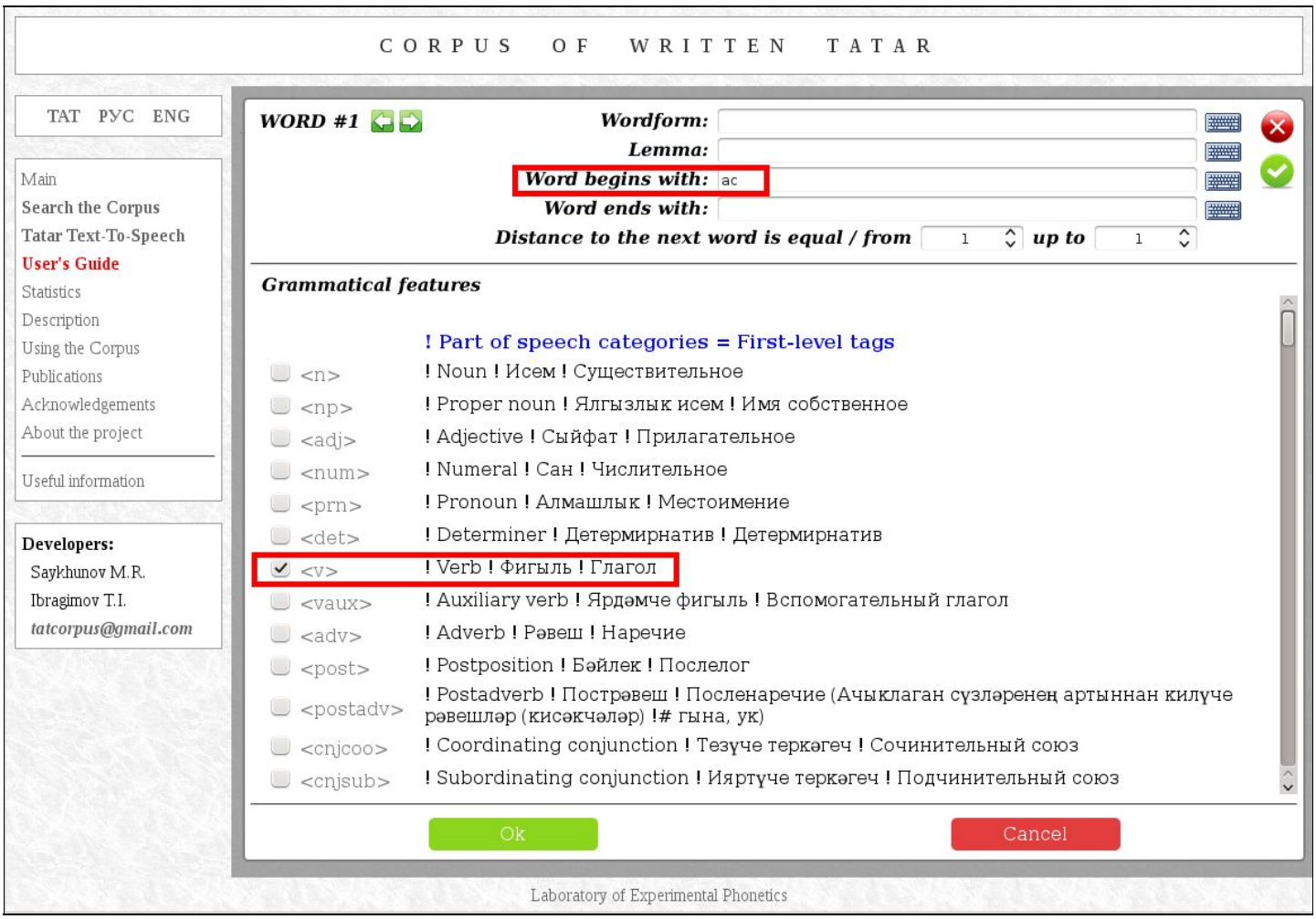
As an example, consider a search for the word form «**алма**» (*noun*). As the first step, enter the word «**алма**» into the word form field, and then, select the appropriate line in the tag list.



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**».

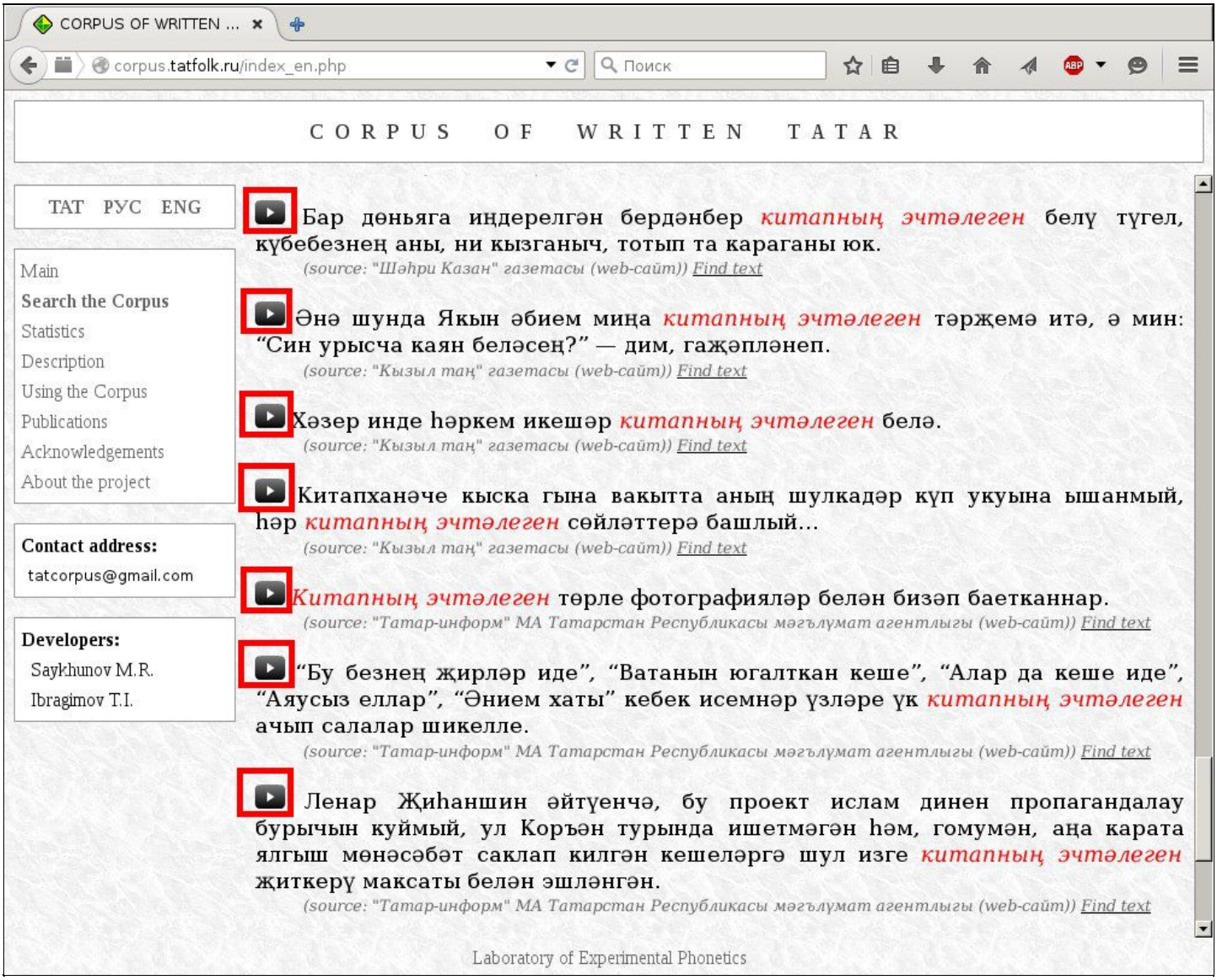


As yet another example, consider the task of finding verbs (**<v>**) starting with a given combination of letters («**ac***»). This request should be set in the graphical mode as follows:



How to listen to the sentences found in the Corpus?

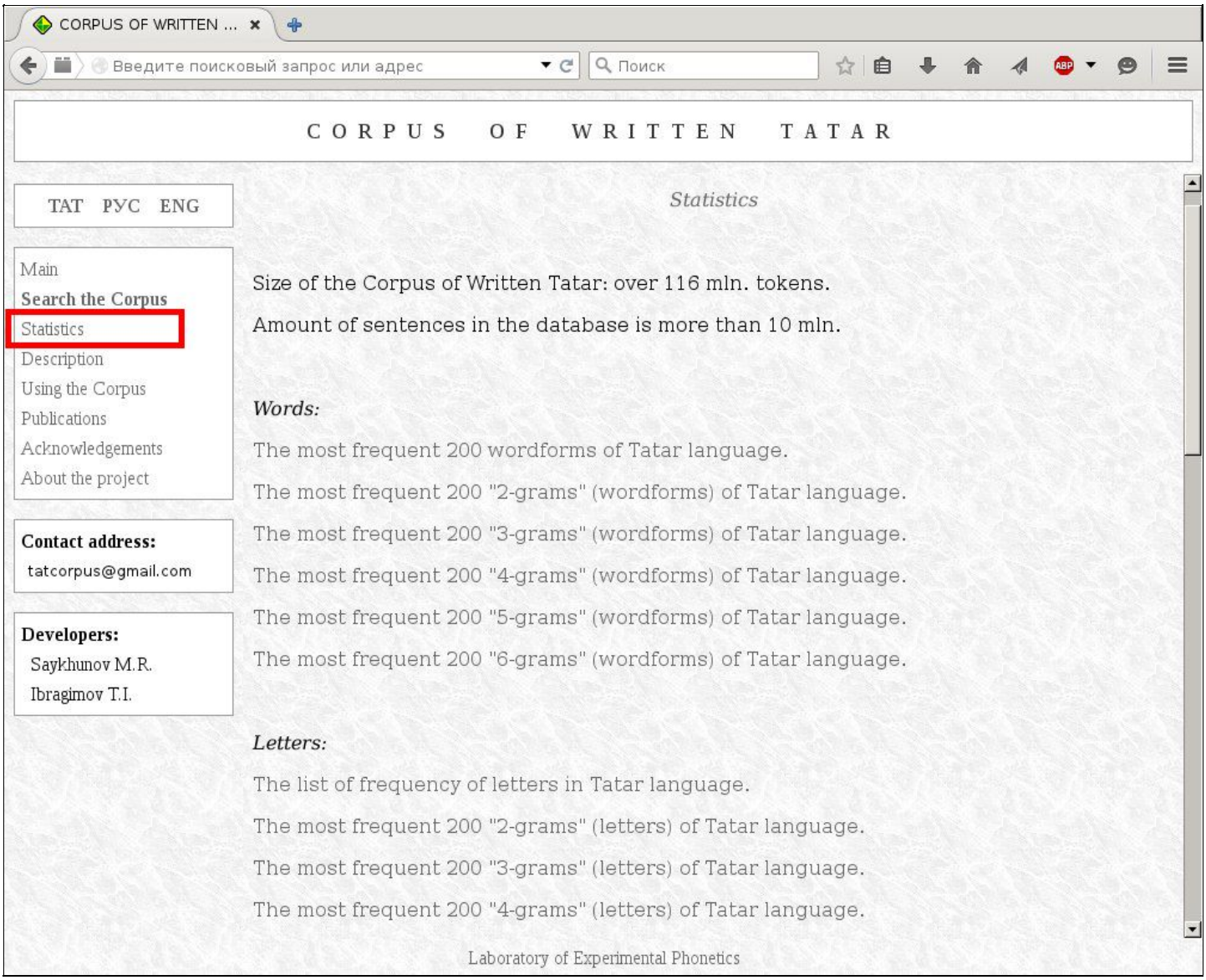
The Corpus of Written Tatar offers the user a unique opportunity to listen to the sentences found in a search! To do this, just click the special button placed to the left of each sentence.



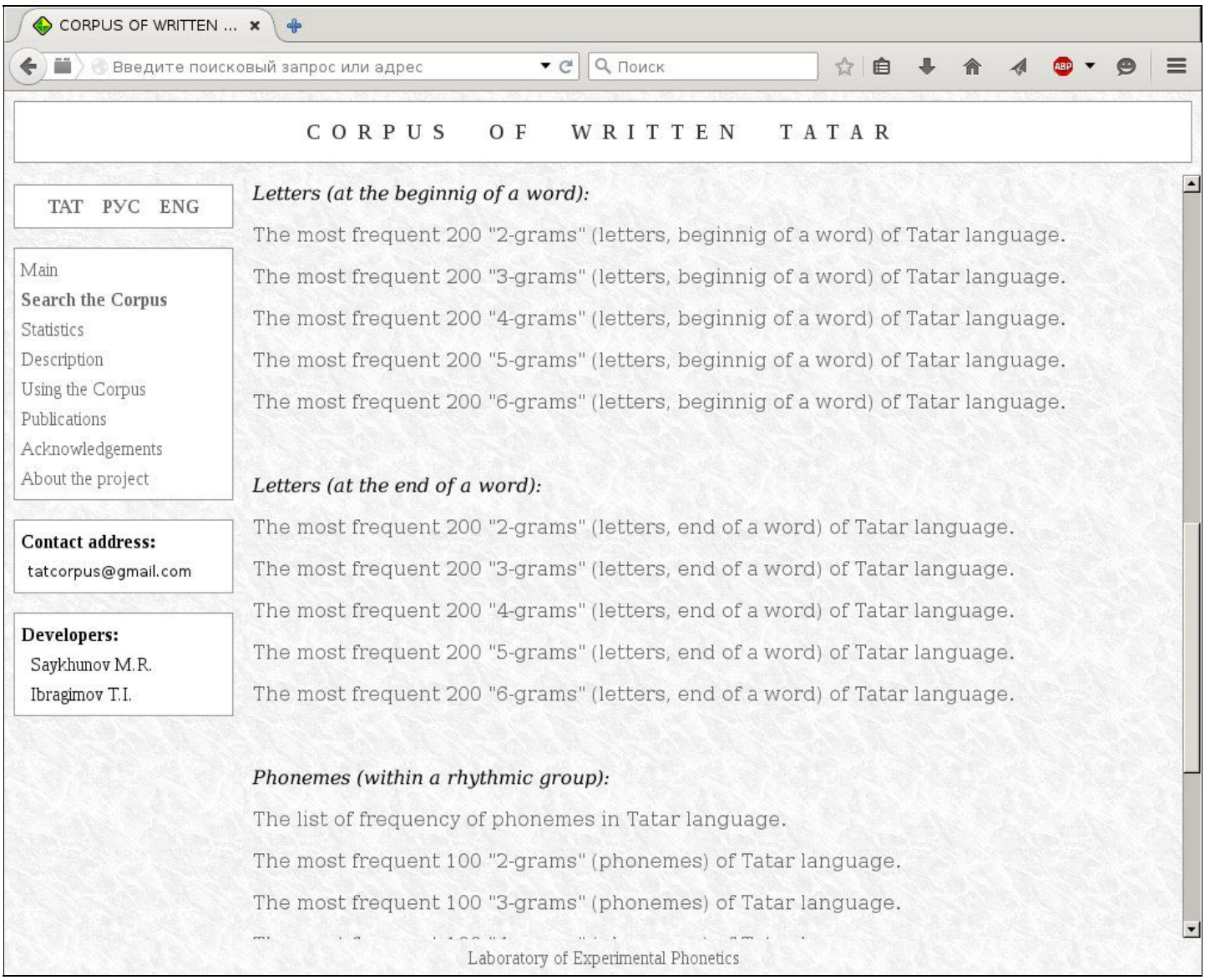
Additional statistical materials!

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:

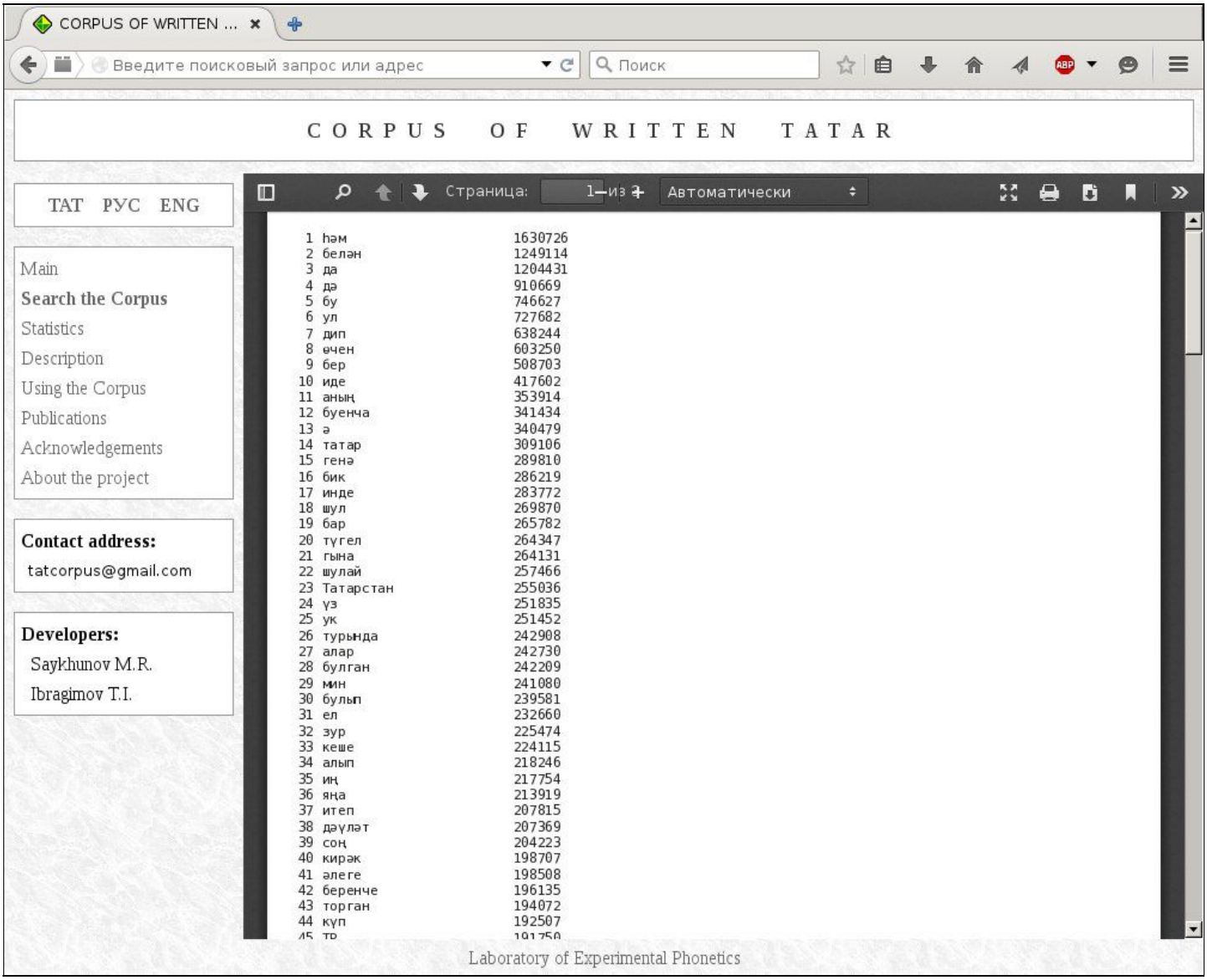
- A list of the most frequent word forms of the Tatar language.
- A list of the most frequent collocations consisting of 2, 3, 4, 5 and 6 words.
- Frequency lists of letters and their combinations.



- 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.

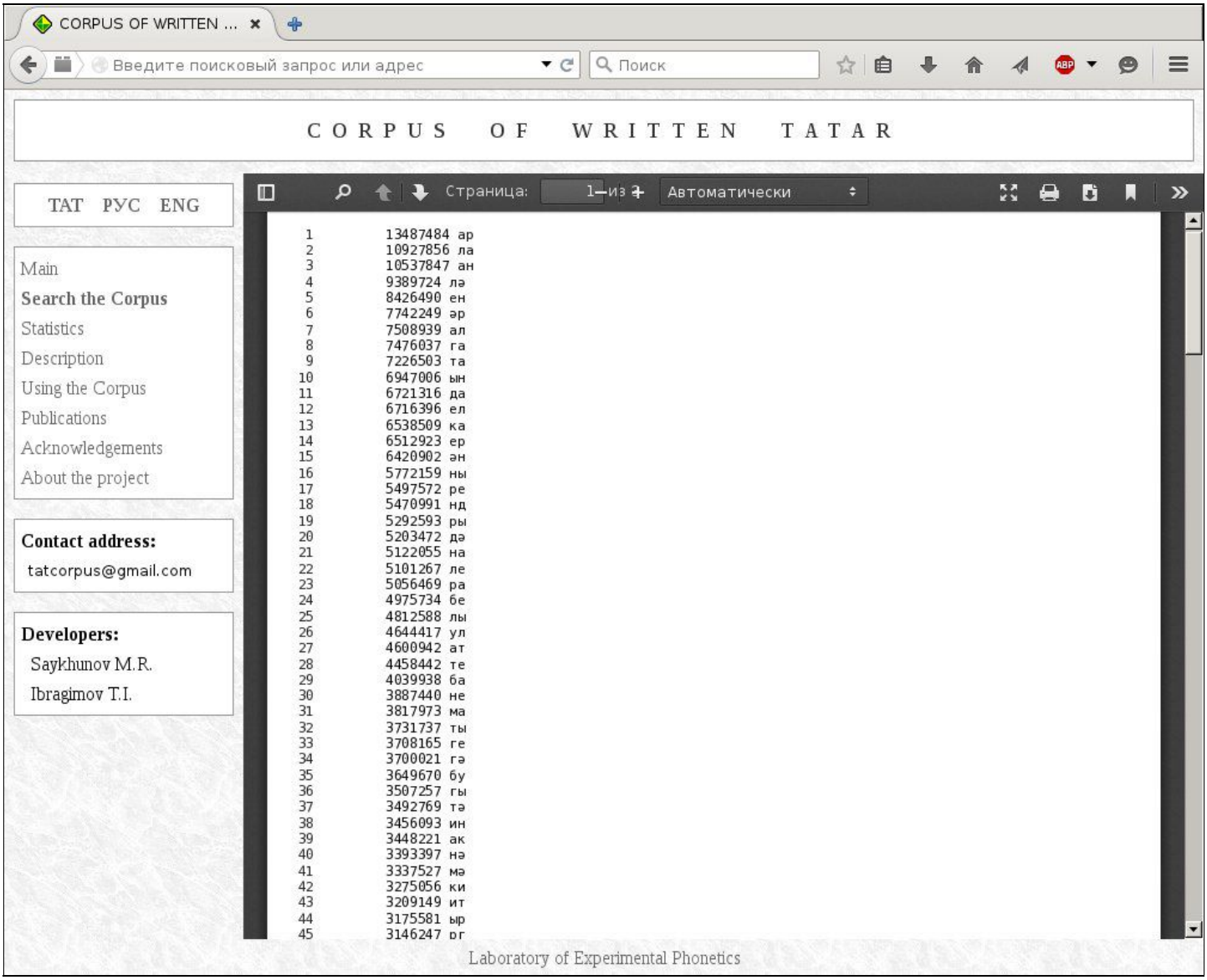


For example, the list of the 5000 most frequent word forms of Tatar language is as follows:



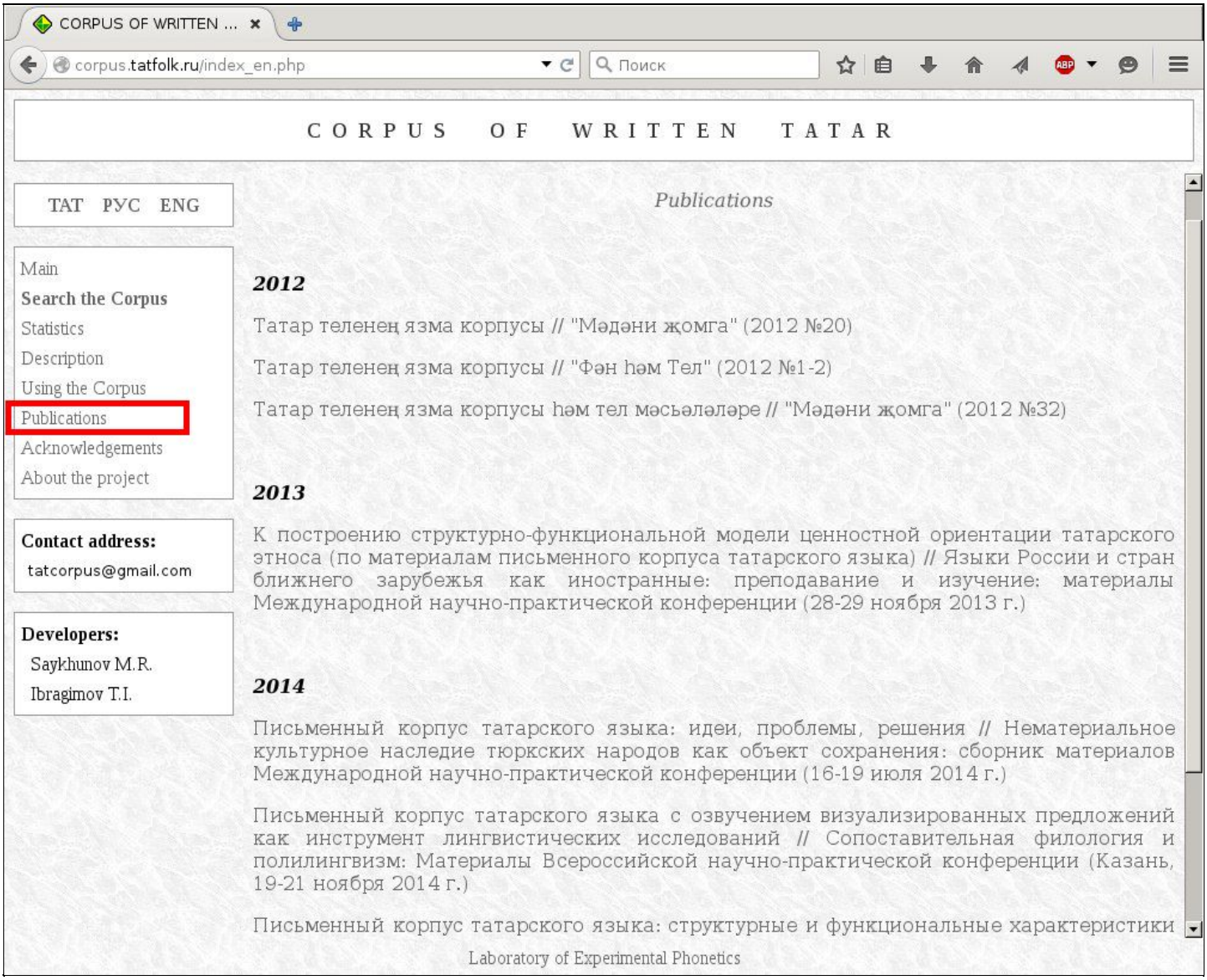
The first column of the list shows the rank of the word form, the second gives the word form itself, and in the third column one can see the number of occurrences in the Corpus.

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



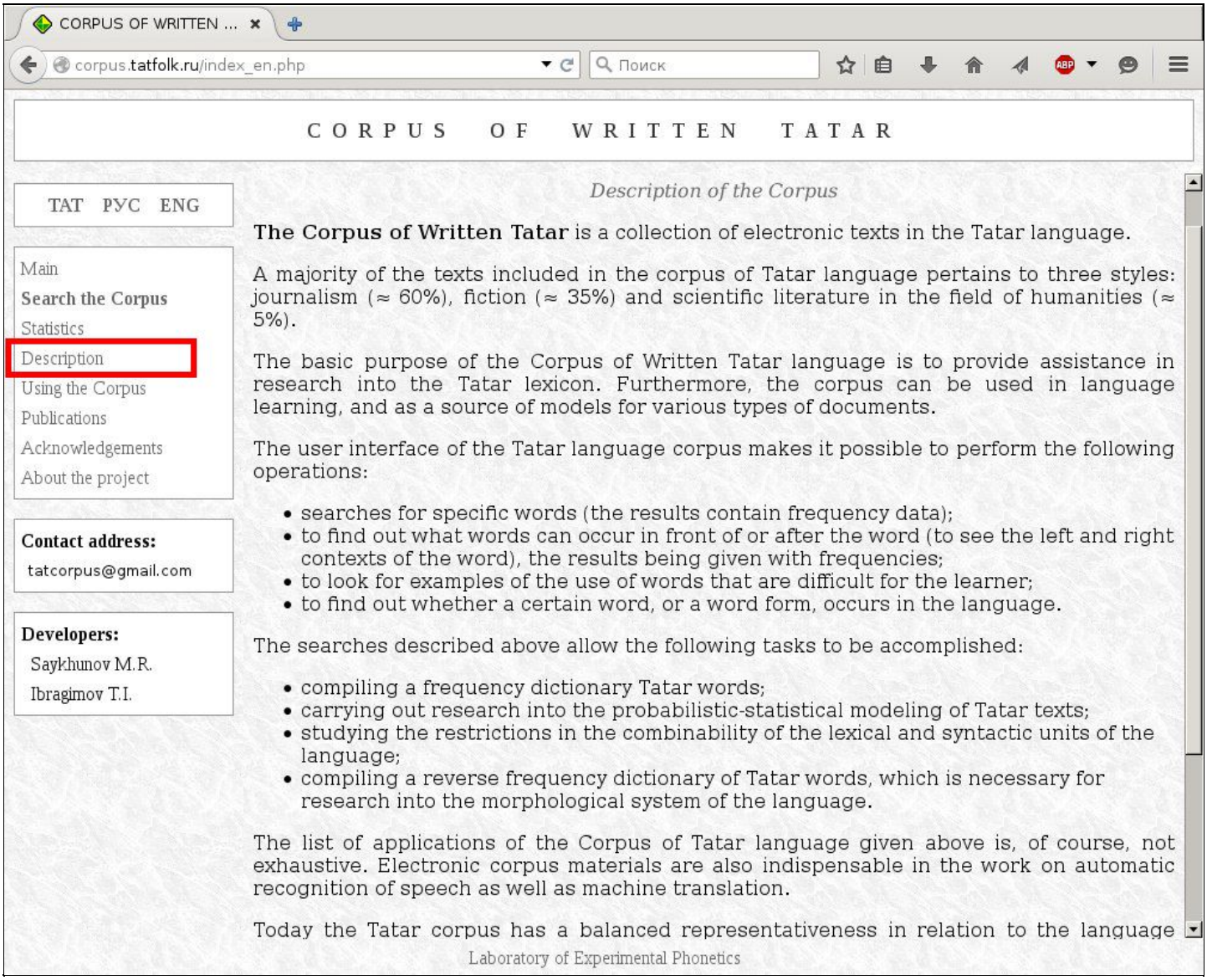
Publications

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



Description

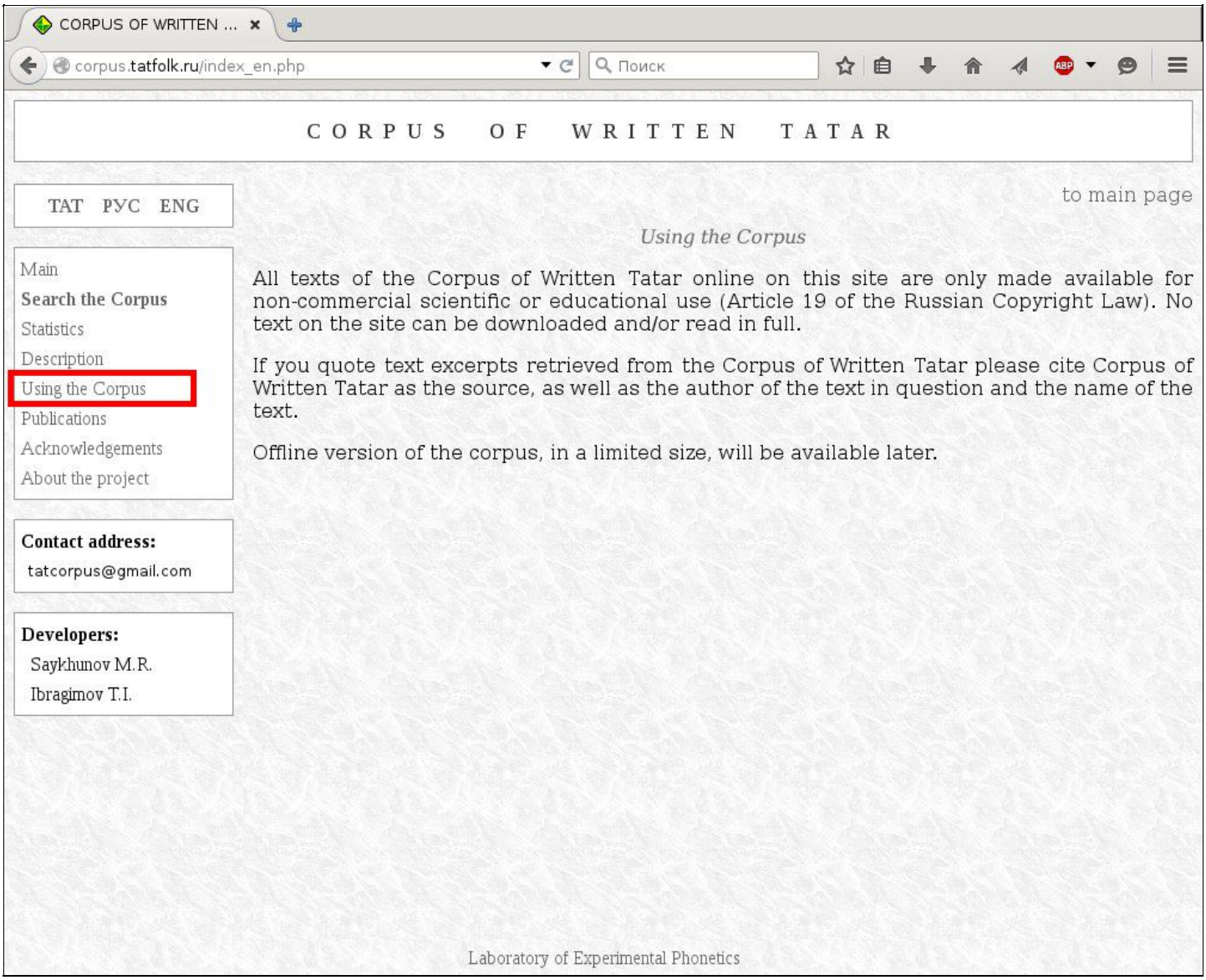
The website's "**Description**" section provides brief information on the structure and functionalities of the Corpus of Written Tatar:



Use and citation

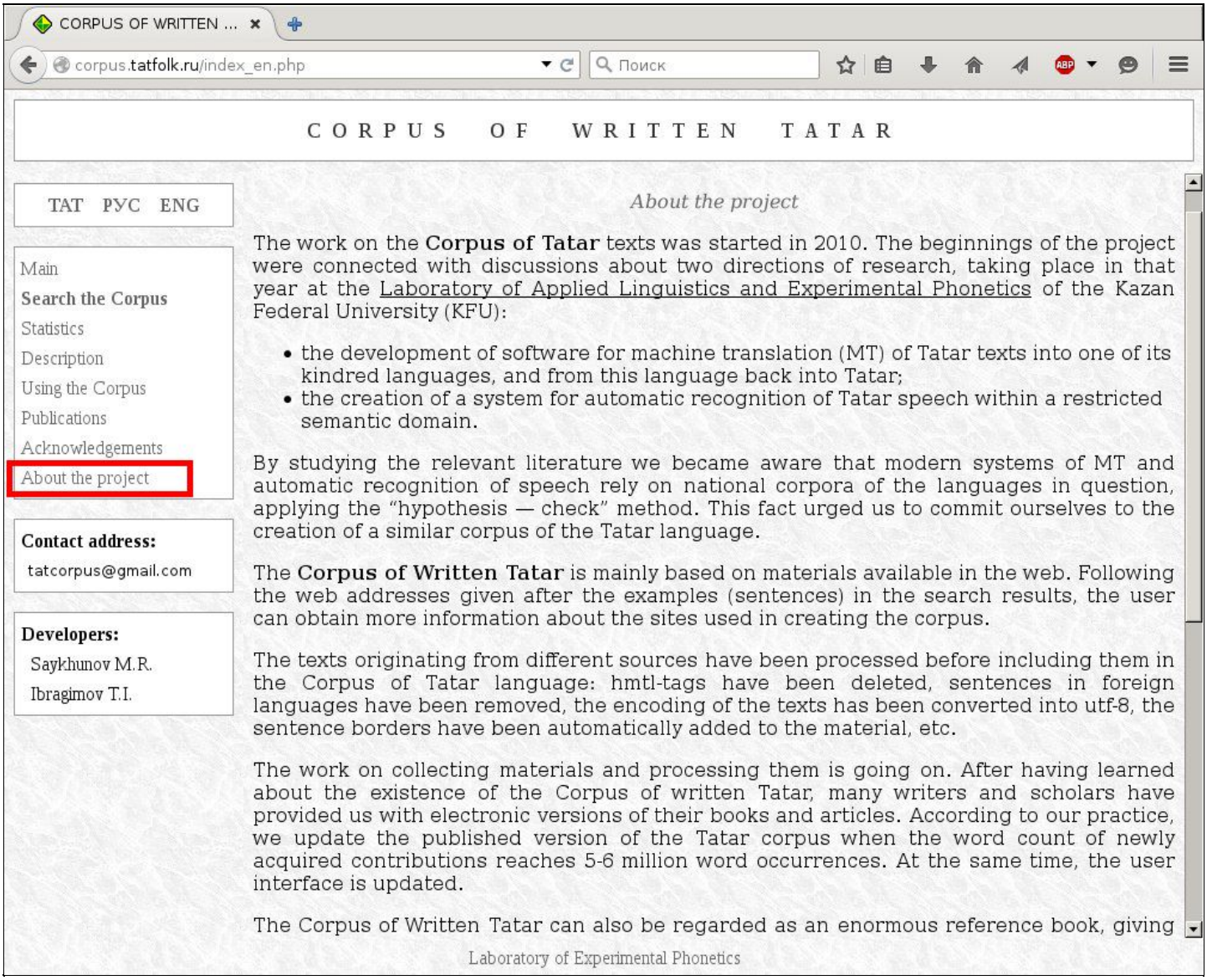
The section "**Using the Corpus**" informs the user on the legal regulations concerning the use of the Corpus, and it also gives recommendations for citing the texts and referencing to the materials of the Corpus, as well as the Corpus itself.

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.



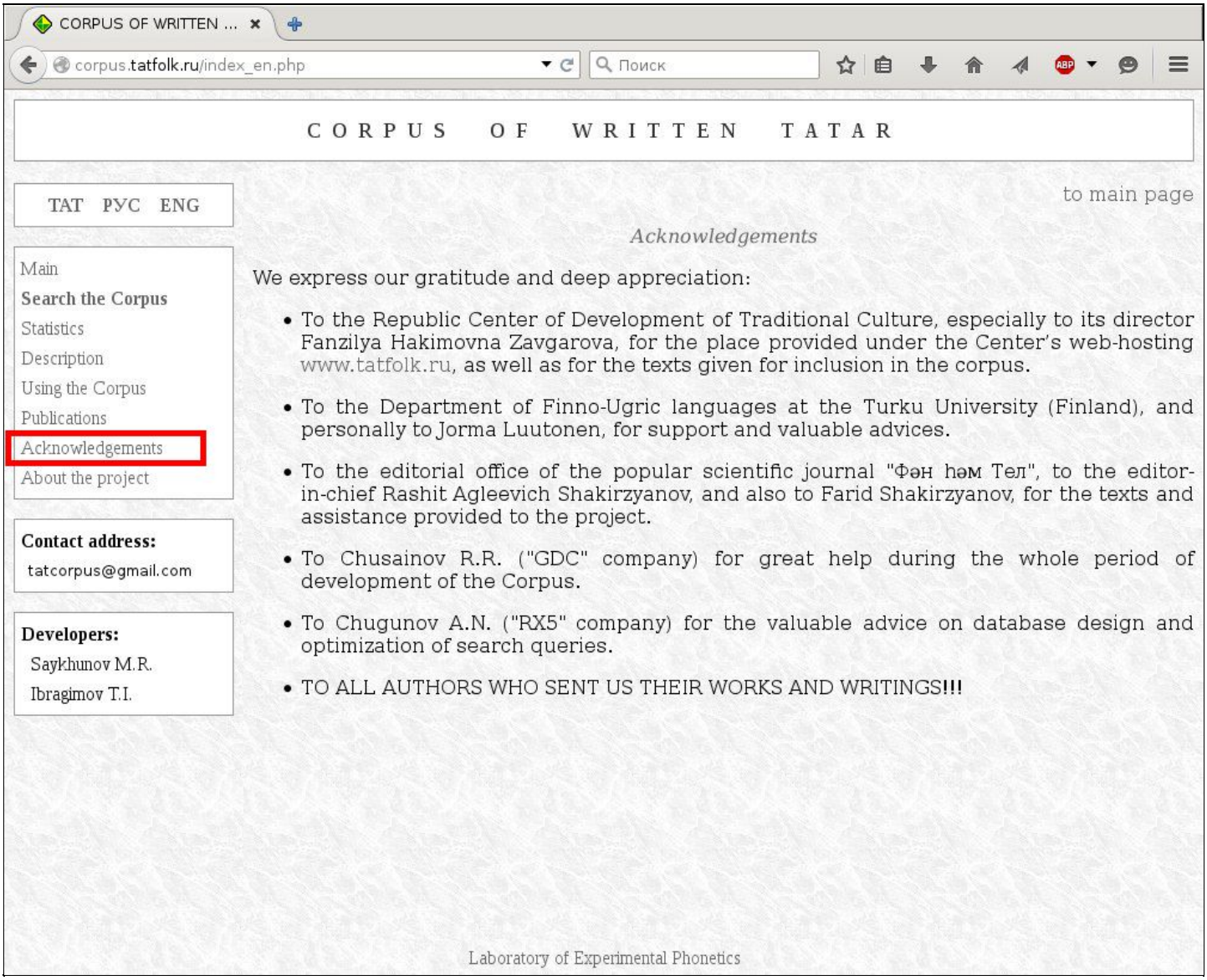
About the project

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



Acknowledgments

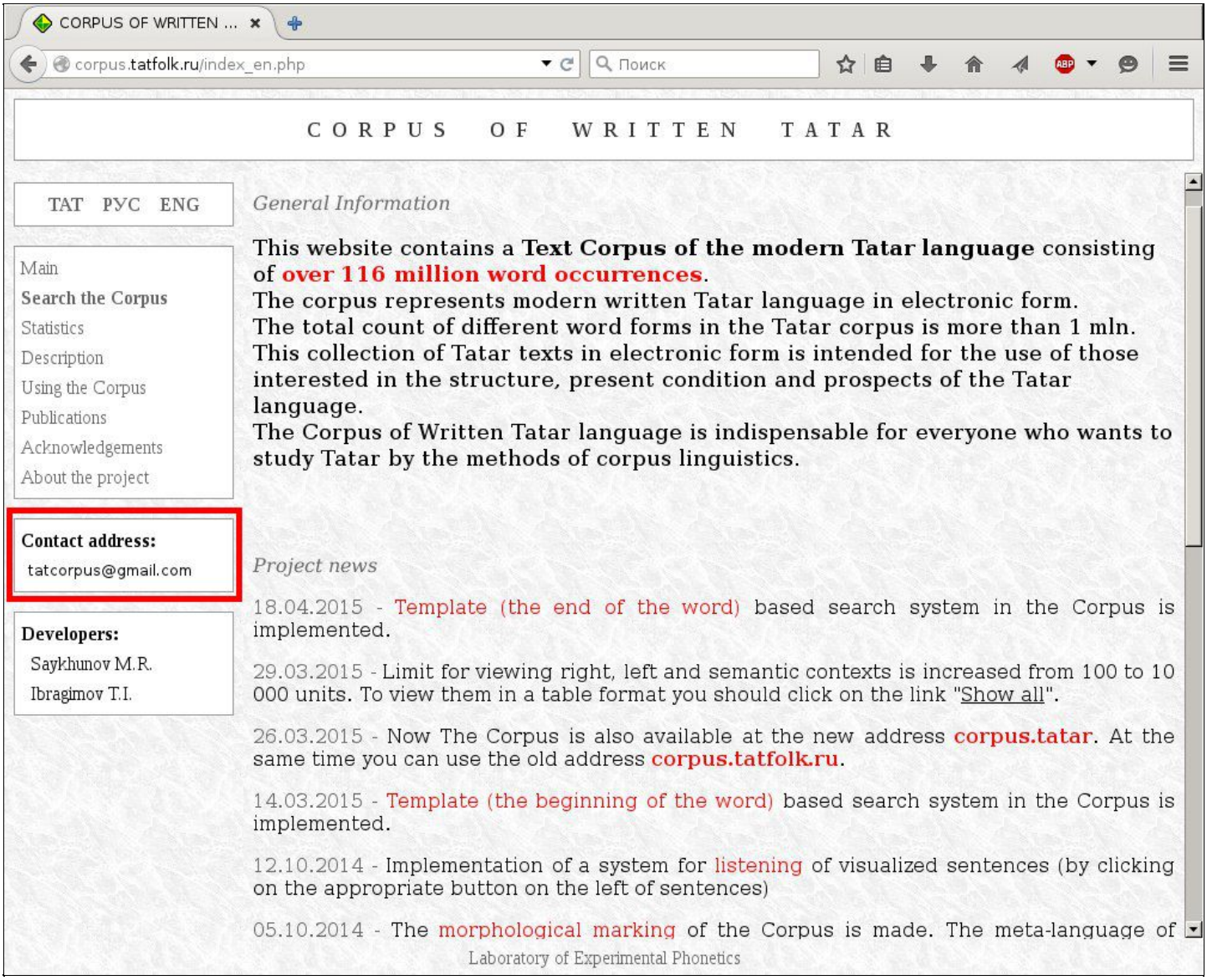
In the section "**Acknowledgments**", the creators of the Corpus wish to express their appreciation to people and organizations that provided great assistance in the work:



Contacts

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

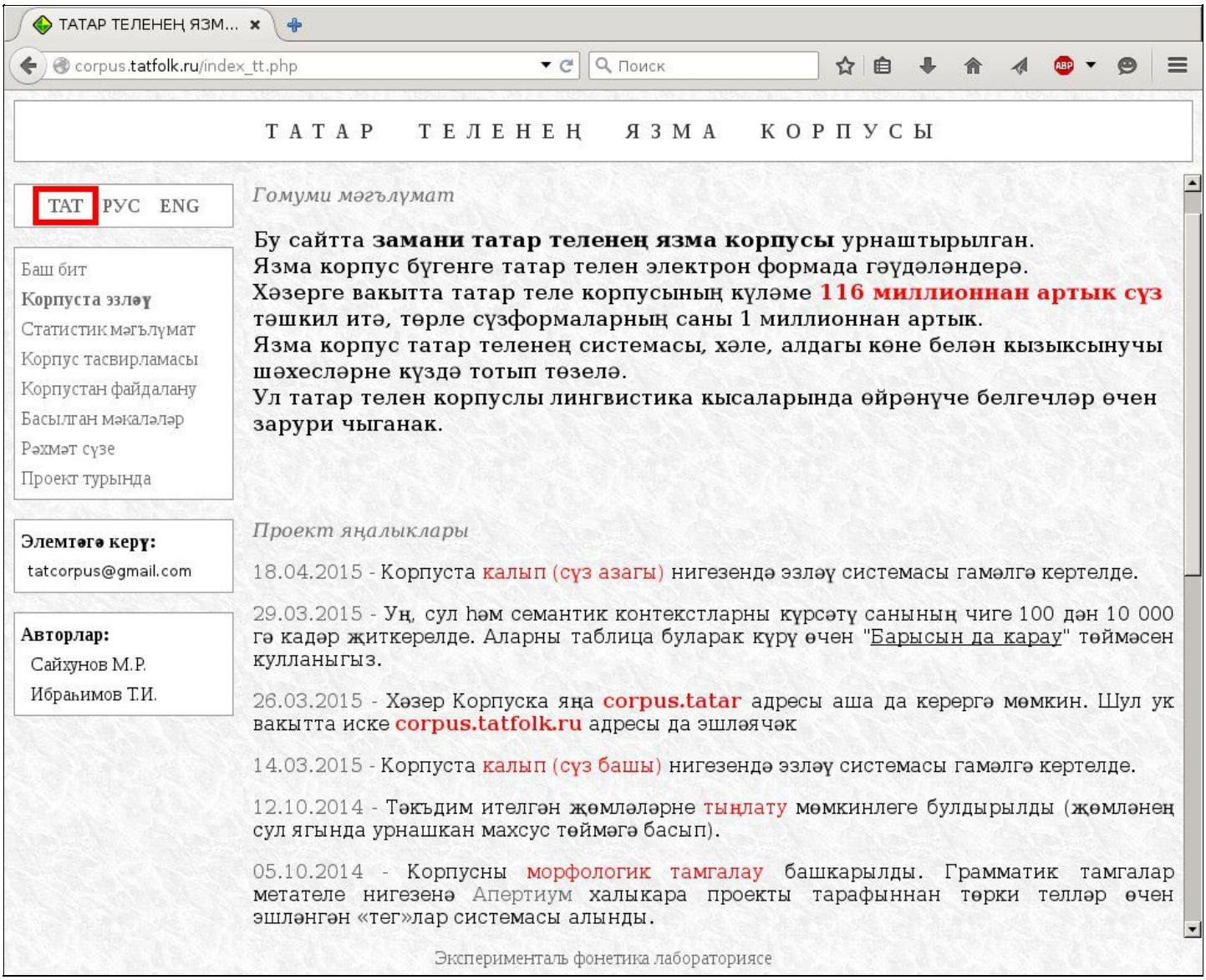
Please contact us if you have any questions or suggestions concerning the Corpus and its use!



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):



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