

# Корпус цыганского языка

ресурс: <http://web-corpora.net/RomaniCorpus/>

Выполнили:

Василиса Кутузова (nemirw@gmail.com)

Елисей Сажин (sazhin.elisey@gmail.com)

Алена Утробина (kladiki@mail.ru)

Мария Чудновская (mlc999@list.ru)

Проект по КИЛИ

## Введение

В этой работе мы рассмотрим Корпус цыганского языка (Russian Romani Corpus), который включает тексты, изданные в СССР в 1920-1930 гг. На данный момент корпус находится в разработке, однако в нём уже зафиксированы 720 тыс. словоупотреблений. Корпус имеет несколько слоев разметки: метатекстовую (библиографическую) и грамматическую.

## Дизайн

В целом, дизайн корпуса довольно приятный. На первый взгляд может показаться, что он слишком минималистичен. Например, главная страница достаточно блеклая: светло-серый заголовок «Корпус цыганского языка» теряется на общем фоне. Также странно выглядит «шапка» - крайне миниатюрная кнопка переключения языков и ссылка для возвращения на главную страницу. На наш взгляд, хорошим решением было бы переместить несколько неприметную строку точного и неточного поиска, находящуюся внизу страницы, в шапку.

Однако данное оформление смотрится вполне выигрышно при непосредственной работе с самим корпусом. Не считая флага в заголовке, цветовое решение выдержано в оранжевых оттенках, причем удачно. Окно ввода форм/лемм/слов для перевода крайне интуитивно - важные части разделены цветом, но все равно сохраняют целостность. Работать с информацией по запросу тоже приятно - создатели постарались и не «засорили» страницу яркими цветами и большим количеством всплывающих окон. Вместо этого они расставили акценты, пользуясь разными шрифтами, жирностью текста и т.д. Таким образом, текст выдачи размечен, но не перегружен, что позволяет пользователю сосредоточиться на работе.

Пример:

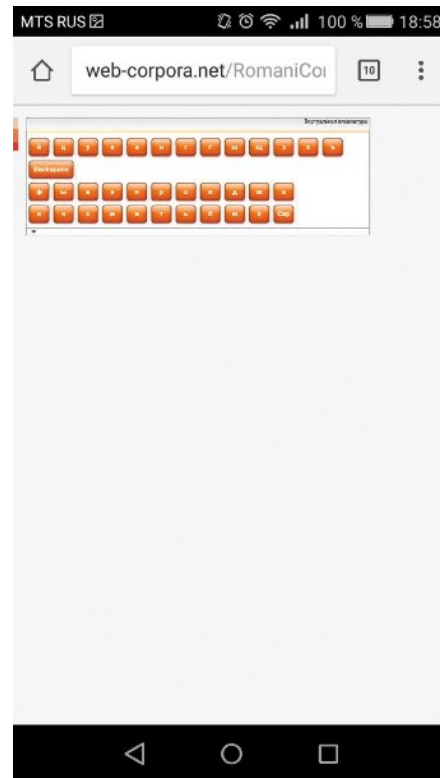
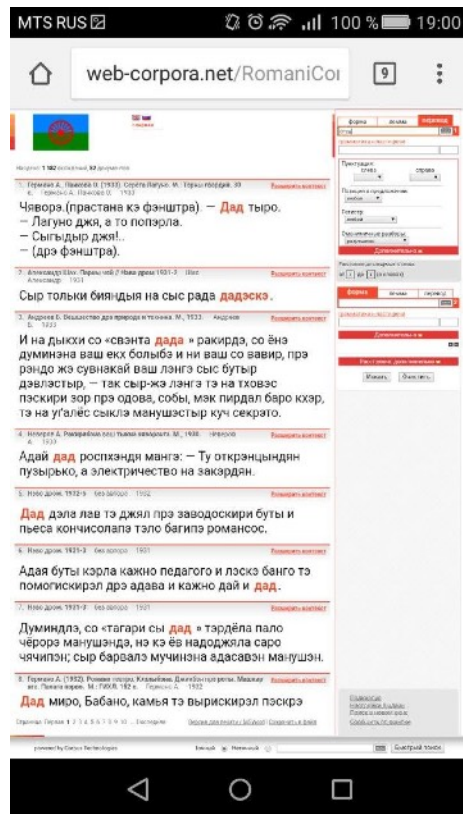
- количество вхождений и документов (именно числа) подсвечены жирным;
- информация об источнике (автор, название, издательство, год) выделена темно-серым цветом, что не отвлекает от контекста, но и не теряется за счет жирного выделения;
- искомое слово/лемма выделены оранжевым.

The screenshot displays the Russian Romani Corpus web application. The main content area shows search results for the word "дадз" (dadz). The results are listed in a table with columns for the source (author, title, publisher, year) and the word's usage in context. The word "дадз" is highlighted in orange in the text snippets. The sidebar on the right contains navigation links (форма, лемма, перевод) and search filters (грамматика и части речи, Дополнительно). The bottom of the page includes a footer with the Corpus Technologies logo and a search bar.

powered by Corpus Technologies

Отдельно стоит отметить мобильную версию сайта. Хотя он поддерживается на мобильном устройстве и передается без искажений, корпус остается крайне неудобным в работе на телефоне. Например, виртуальная клавиатура открывается новой вкладкой на весь экран - пользователь не имеет возможности в режиме реального времени оценить набранный текст.

На небольшом экране страница выглядит чересчур сжатой и воспринимается гораздо хуже потому, что значительную часть экрана всегда занимает правое окно ввода и поиска с дополнительными настройками.

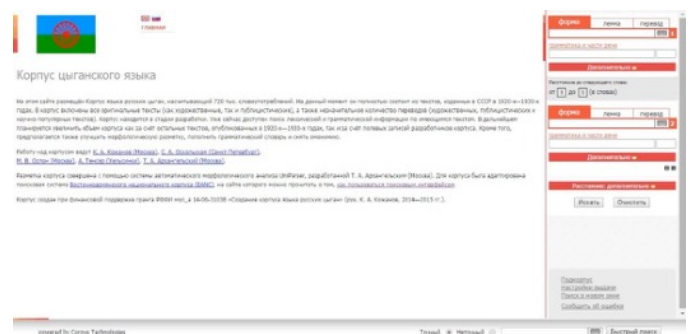
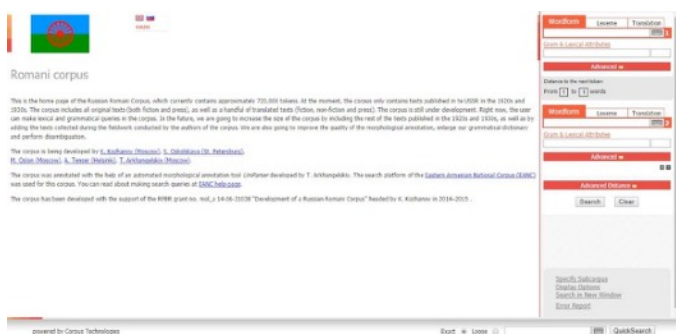


Итак, из-за

своего дизайна корпус кажется сильно недоработанным, но является крайне практичным и удобным в работе, то есть новичку легко в нем ориентироваться во многом благодаря создавшейся «пустоте» (или ощущению «пустоты»/незаполненности). Относительно красоты - сложно судить, поскольку мы бы, например, постарались сделать оформление более лаконичным и сглаженным, добавив подложки к заголовкам на главной странице и разделив информацию на смысловые куски. Так можно будет придать более завершённый вид странице. Также, наверное, резонно будет доделать раздел настроек (нижняя правая часть страницы).

## Ресурс глазами новичка

Корпус языка русских цыган достаточно просто находится в интернете: при поиске по запросам «корпус цыганского языка» или “romani corpora”, “romani corpus” в Google ссылка на соответствующую версию ресурса - русско- или же англоязычную, в зависимости от запроса, - будет самой первой в результатах выдачи. Так выглядит главная страница корпуса:



Переключение между англо- и русскоязычными версиями интуитивно понятно и осуществляется путем нажатия на соответствующую кнопку-флажок вверху страницы. Настройки поиска находятся здесь же, в правой части главной страницы; кроме того, не составляет труда найти ссылки на создание подкорпуса и настройку выдачи результатов (открываются в отдельном окне), форму быстрого поиска и специальную кнопку, позволяющую начать поиск в новой вкладке, что, безусловно, удобно при необходимости проведения каких-либо параллельных исследований. Таким образом, на первый взгляд интерфейс корпуса выглядит понятным, не вызывает резкого отторжения или непонимания.

## Помощь пользователю

Отдельный раздел помощи пользователю на сайте отсутствует как таковой. Тем не менее, на сайте присутствует ссылка на подробную инструкцию к поиску по Восточно-армянскому национальному корпусу, адаптацией поисковой системы которого пользуется и Корпус цыганского языка.

Инструкция к ВАНК значительно упрощает поиск по корпусу примерами и скриншотами, но, к сожалению, здесь существуют некоторые подводные камни. Дело в том, что не все функции поиска, работающие в Восточно-армянском национальном корпусе, применимы к КЦЯ. Так, в рассматриваемом нами ресурсе отсутствует функция перевода словосочетания. Если мы попробуем действовать по инструкции пользователя ВАНК, то столкнемся с ошибкой:

The screenshot shows the VANK search interface. At the top, there are tabs for 'форма' (form), 'лемма' (lemma), and 'перевод' (translation). The 'перевод' tab is selected, and the search input contains the text '"give up"'. Below the input, there is a red button labeled '1'. Underneath, there is a section for 'грамматика и части речи' (grammar and parts of speech) with a dropdown menu. Below that is a red button labeled 'Дополнительно' (Additional). Further down, there is a section for 'Расстояние до следующего слова:' (Distance to the next word:) with a dropdown menu. Below that is a red button labeled '2'. At the bottom, there are two buttons: 'Искать' (Search) and 'Очистить' (Clear).

Искомый элемент запроса не найден.  
Перевод '"give

В общем и целом, набор функций Корпуса цыганского языка приближен к «стандартному» для любого корпуса набора: существует поиск по словоформе, по лемме, по грамматическим признакам, возможен поиск сочетаний из нескольких слов, пользователь также может скачать нужную ему выборку и так далее. В частности, интерфейс в чем-то схож с интерфейсом Национального корпуса русского языка и корпусами других языков на базе web-corpora.net. Из особенностей можно отметить виртуальную клавиатуру со специальными символами, особенно полезную для тех, у кого отсутствует кириллическая раскладка.

Суммируя, можно сказать следующее: если к ресурсу обратится пользователь, уже имеющий опыт работы с различными корпусами, то весьма вероятно, что поиск по КЦЯ будет осуществляться им почти интуитивно, и отсутствие раздела помощи или неточная инструкция практически не осложнят его работу. Если же пользователь никогда не

работал со схожими ресурсами, велик риск возникновения непонимания механизмов работы поиска.


## Продвинутый функционал

Сделаем следующий запрос. Посмотрим на частотность леммы “гож” (красивый) в художественных (результат 1) и нехудожественных (результат 2) текстах. К сожалению, задать какие-либо еще критерии (допустим, чтобы мы увидели исключительно примеры в единственном числе и женского рода) не удастся, поскольку корпус относительно невелик.

The screenshot displays the search results for the word "гож" in the Corpus Technologies database. The interface includes a header with the Corpus Technologies logo and a navigation bar. The search results are listed in a table with columns for the search term, the source text, the author, and the year. The results are sorted by frequency, with the top result being "гож" from the text "Адаптация" by М. Безлюдов, 1936. The search results are displayed in a table with columns for the search term, the source text, the author, and the year. The results are sorted by frequency, with the top result being "гож" from the text "Адаптация" by М. Безлюдов, 1936.

Результат	Текст	Автор	Год
1	Адаптация	М. Безлюдов	1936
2	Адаптация	М. Безлюдов	1936
3	Адаптация	М. Безлюдов	1936
4	Адаптация	М. Безлюдов	1936
5	Адаптация	М. Безлюдов	1936
6	Адаптация	М. Безлюдов	1936
7	Адаптация	М. Безлюдов	1936

результат 1



ГЛАВНАЯ

Найдено: **6** экземпляров, **2** документов

Размер подкорпуса: **0.66%** от общего объёма корпуса

1. — — —

Расширить контекст

Ту ака, со, мро чяваро, — ракирла дай, — ма шундём, со дро клубо исы экс романы чай, ракирна драван лылвари и **гожо**.

2. — — —

Расширить контекст

Адала козлыбана сыкада аменгэ со адала чявэ на джяна пиро дром амада дадан, ёна лана сыклякирда манушэнца и ангил гонгиро дром лапа бугло и **гожо**.

3. — — —

Расширить контекст

На боюк, шыл, вышутимэ чяворэнгирэ муй и одова, со ёно нангэ, мзалап и со пашыл ланса чявэ бутитконэн тато и гожэс урида и муй здорова и **гожа** сыр паба, — заухтылла ромэн са бутыр и бутыр ко ново джибон.

4. — — —

Расширить контекст

Адай сыс здорова, бахтало и **гожо** сыр коамитко дыкас козлыбан.

5. — — —

Расширить контекст

Митроскэ исыс 20 барш, **гожо** про муй и саро исыс дрэ дядэстэ кжоро...

6. — — —

Расширить контекст

Лёля фронтэс адыхои проз Ганатэ, проракирда: —!л — Драван **гожо**...

форма

ЛОНЧОЗ

перевод

гож

граничати и части речи

Дополнительно

Расстояние: дополнительно

Искать

Очистить

Подкорпус:

Настройки выдачи

Поиск в новом окне

Сообщить об ошибке

Страница: Первая 1 Последняя

Воскря для печати / MSWord | Сохранить в файл

powered by Corpus Technologies

Точный

Неточный


Быстрый поиск


## результат 2

Корпус показывает, что в художественных текстах мы встретим эту форму чаще. Примеры данного запроса составляют чуть больше 42% от объема корпуса, в то время как результаты второго запроса - меньше 1%.

Попробуем сделать запрос, позволяющий определить частотность использования существительных и местоимений в качестве подлежащего в конструкциях типа “подлежащее-сказуемое-прямое дополнение” среди прозаических и поэтических произведений. Для этого зададим в двух параллельных вкладках подкорпус с прозой и выполним поиск для подлежащего-существительного:







главная

Найдено: **81** вхождений, **37** документов  
Размер подкорпуса: **94.23%** от общего объема корпуса

1. Германов А. (1932). Романо театро. Кэзлыбэна. Джиибэн прэ роты. Машкир яга. Палага пэрво. М.: ГИХЛ. 152 с. Германов А. 1932	<a href="#">Расширить контекст</a>
На дэ годла!...	
2. Германов А. (1935). Ганка Чямба и ваврэ роспэныбэна. М.: Художественная литература. 102 с. Германов А. 1935	<a href="#">Расширить контекст</a>
Васта тринскирдэпо холятыр.	
3. Нэво дром. 1931-8 без автора 1931	<a href="#">Расширить контекст</a>
Пионеры гына кэра.	
4. Ляшко Н. Роспэныбэн ваш случаё. М., 1935. Ляшко Н. 1935	<a href="#">Расширить контекст</a>
Этако кэсдэ ягэнца.	
5. Нэво дром. 1932-2-3 без автора 1932	<a href="#">Расширить контекст</a>
Хасэ отлыджыла колхозостыр.	
6. А.Г. Советско сэидо // Романы зоря 1929-2 А. Г. 1929	<a href="#">Расширить контекст</a>
Табуро чэрдэ грэн.	
7. Германов А. (1932). Романо театро. Кэзлыбэна. Джиибэн прэ роты. Машкир яга. Палага пэрво. М.: ГИХЛ. 152 с. Германов А. 1932	<a href="#">Расширить контекст</a>
2-ро парно офицэро.	
8. Нэво дром. 1931-4-5 без автора 1931	<a href="#">Расширить контекст</a>
Строна барыёла заводэнца.	

Страница: Первая 1 2 3 4 5 6 7 8 9 ... Последняя

Версия для печати / MSWord | Сохранить в файл

powered by Corpus Technologies

Точный ☒ Неточный ☐

Быстрый поиск.

Пунктуация: слева  справа

Позиция в предложении: в начале

Регистр: любой

Омонимичные разборы: разрешены

Дополнительно

Расстояние до следующего слова: от 1 до 1 (в словах)

форма  лемма  перевод

грамматика и части речи

Пунктуация: слева  справа

Позиция в предложении: в середине

Регистр: любой

Омонимичные разборы: разрешены

Дополнительно

Расстояние до следующего слова: от 1 до 1 (в словах)


форма  лемма  перевод


грамматика и части речи

Пунктуация: слева  справа

Позиция в предложении: в конце

...и подлежащего-местоимения:





главная

Найдено: **60** вхождений, **28** документов  
Размер подкорпуса: **94.23%** от общего объема корпуса

1. Светлов Л. (1938). Ром Хвасю. М.: Художественная литература. 136 с. Светлов Л. 1938	<a href="#">Расширить контекст</a>
Миро лав састэр.	
2. А.С. Пушкин. Капитаноскири чай. Москва: Художественная литература. 1938 Пушкин А. С. 1938	<a href="#">Расширить контекст</a>
Ёно сы чёра.	
3. Германов А. (1932). Романо театро. Кэзлыбэна. Джиибэн прэ роты. Машкир яга. Палага пэрво. М.: ГИХЛ. 152 с. Германов А. 1932	<a href="#">Расширить контекст</a>
Сарэ дыкэна э-рувэнца	
4. Л. Толстой. Трин рыча. Москва 1937 Ленинград Толстой Л. Н. 1937	<a href="#">Расширить контекст</a>
Вавир сым дай.	
5. Светлов Л. (1938). Ром Хвасю. М.: Художественная литература. 136 с. Светлов Л. 1938	<a href="#">Расширить контекст</a>
Адава сы парамыся.	
6. Германов А. (1932). Романо театро. Кэзлыбэна. Джиибэн прэ роты. Машкир яга. Палага пэрво. М.: ГИХЛ. 152 с. Германов А. 1932	<a href="#">Расширить контекст</a>
Саво мэ кулако?	
7. Германов А. (1933). Лэс кэрдэ рувэса и ваврэ роспэныбэна. М.-Л.: Художественная литература. 119 с. Германов А. 1933	<a href="#">Расширить контекст</a>
Токо-со багандэ гилы.	
8. Германов А. (1935). Ганка Чямба и ваврэ роспэныбэна. М.: Художественная литература. 102 с. Германов А. 1935	<a href="#">Расширить контекст</a>
Вавир думиндя Ненила.	

Страница: Первая 1 2 3 4 5 6 Последняя

Версия для печати / MSWord | Сохранить в файл

powered by Corpus Technologies

Точный ☒ Неточный ☐

Быстрый поиск.

Пунктуация: слева  справа

Позиция в предложении: в начале

Регистр: любой

Омонимичные разборы: разрешены

Дополнительно

Расстояние до следующего слова: от 1 до 1 (в словах)

форма  лемма  перевод

грамматика и части речи

Пунктуация: слева  справа

Позиция в предложении: в середине

Регистр: любой

Омонимичные разборы: разрешены

Дополнительно

Расстояние до следующего слова: от 1 до 1 (в словах)

форма  лемма  перевод

грамматика и части речи

Пунктуация: слева  справа

Позиция в предложении: в конце

Далее сделаем то же самое с подкорпусом поэзии. В результате получим следующие данные:

	N, кол-во вхождений	PRON, кол-во вхождений
Проза	81	60
Поэзия	8	0

Отсюда можно сделать вывод, что в предложениях типа “подлежащее-сказуемое-прямое дополнение” в роли подлежащего существительное выступает относительно чаще, чем местоимение.

Проведем критическую оценку функциональности ресурса:

Что позволяет найти ресурс?	Что не позволяет найти ресурс?	Комментарии
возможен перевод с русского языка	имеются недочеты: не все слова присутствуют	рис. 1.1, 1.2 белый цвет есть, а черный - нет
есть перевод с английского	выполнен не в полной мере	рис. 2.1 загадочным образом отсутствуют самые элементарные слова (mother, father)
есть возможность задавать при поиске омонимичные разборы	имеет функцию WITHOUT_GLOSS, значение которой, к сожалению, нам так и не удалось выяснить	
	не снята омонимия	
возможен поиск по времени издания произведений	базу корпуса составляют тексты 1920-ых - 1930-ых гг.	довольно скудная выборка для корпуса
есть элементы регулирования пунктуации слева и справа от искомого слова	можно выбрать только отсутствие пунктуационных знаков или наличие любого из них (без конкретизации)	

рис. 1.1



Найдено: **596** вхождений, **72** документов

1. Лебедев Н.К. Есхджини машкир дикарендэ. М., 1937.    Лебедев Н.К.    1937    [Расширить контекст](#)

Пало послед-я штарадэша борша папуас удиконэ, со машкир **парнэ** ваэрэ пхувитконэ манушэнда, савэ явэна пэлэ морёстыр, дрэван набут мануша, савэ здэна прэ Маклаэсты.

2. Светлов Л. (1938). Ром Хвасю. М.: Художественная литература. 136 с.    Светлов Л.    1938    [Расширить контекст](#)

Окэ и ачём нэ пхуро, еоджино про **парно** сэвто.

3. Лебедево Н.К. Архангельска Робинзоны. М., 1935.    Лебедево Н.К.    1935    [Расширить контекст](#)

Чечено хулай дра полярно область — адава **парно** рыч, или, сыр лэс ксарна прэ северо ошкүй.

4. Нэво дром. 1931-8    без автора    1931    [Расширить контекст](#)

Льстэ **парнэ** зорэлэ дэнда, прэ упратуны ушт пиро възглы выджяне кэлэ фанцы, о Илько чэстэс кошэлэлэ пхура ромэнда.

5. Л.Н. Толстой. Коли прогыя бало. М.: Художественная литература, 1936    Толстой Л. Н.    1936    [Расширить контекст](#)

Адава сым лакиро дэд, пэосирэ лолэ чамьэнца и **парнэ** вэнсцэнца и бакэнбардэнца.

6. Светлов Л. (1938). Ром Хвасю. М.: Художественная литература. 136 с.    Светлов Л.    1938    [Расширить контекст](#)

Ужэ **парно** о чён г азынлэ учес и обидя, сыр парнэ тхуэкса, сари шатра, нэ Рахиль на пася.

7. Германо А. (1935). Ганка Чямба и ваврэ роспхэныбана. М.: Художественная литература. 102 с.    Германо А.    1935    [Расширить контекст](#)

Прикоардэ дэвлэ хэлэдэс Ягорийэс и припхидэ тэ ухтэл лосю про **парнэ** грэстэ дро Дарыдаскиро таборо.  
— Ягорийэ, угалёв, сосыр Дарыдаскиро ромэ ман на приглёвэ? Ил бут времё ухтя пхуэ Ягорийэ, нэ на могикирдэ тэ розордал ромэн: ёв дро еск форо яэзла, а ёнэ дро ваэир уджяна.

8. Светлов Л. (1938). Ром Хвасю. М.: Художественная литература. 136 с.    Светлов Л.    1938    [Расширить контекст](#)

А со ада ялоа ланиро, сыр подыкола про тутэ, адяко и касин про **парно** сэвто.

Страница: Первая   1   2   3   4   5   6   7   8   9   10   ...   Последняя

Версия для печати / MSWord | Сохранить в файл

powered by Corpus Technologies

Точный ☒ Неточный ☐

форма    лемма    перевод

белый

грамматика и части речи

Дополнительно

Расстояние до следующего слова:  
от  до  (в словах)

форма    лемма    перевод

грамматика и части речи

Дополнительно

Расстояние: дополнительно

Искать

Подкоопус  
Настройкы выдачи  
Поиск в новом окне  
Сообщить об ошибке

рис. 1.2

Искомый элемент запроса не найден.  
Перевод **черный**

powered by Corpus Technologies

Точный ☒ Неточный ☐

форма    лемма    перевод

черный

грамматика и части речи

Дополнительно

Расстояние до следующего слова:  
от  до  (в словах)

форма    лемма    перевод

грамматика и части речи

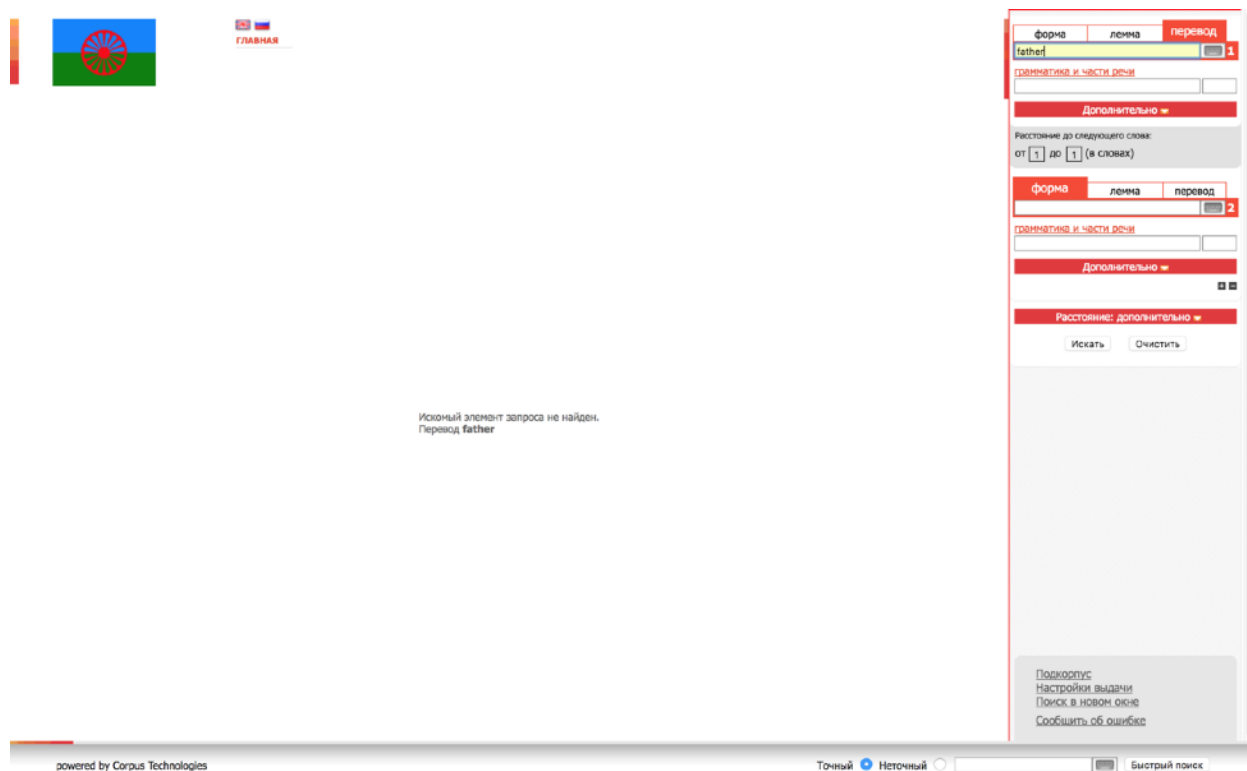
Дополнительно

Расстояние: дополнительно

Искать

Подкоопус  
Настройкы выдачи  
Поиск в новом окне  
Сообщить об ошибке

рис. 2.1



## Выводы

Исходя из приведенного описания можем заключить, что Корпус цыганского языка интуитивно понятен и прост в применении, хоть и обладает недостаточно широким функционалом. По сравнению с НКРЯ, данный корпус имеет более скудные возможности для работы в оффлайне, меньшее количество слоёв разметки и менее приятный для глаз интерфейс. Корпус позволяет совершать простые запросы, но вариативность во многом ограничивается не слишком большим количеством и разнообразием текстов. В дизайне присутствуют несколько недочетов, однако они не затрудняют работу с корпусом.