



# Корпус цыганского языка

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

Выполнили:

Василиса Кутузова (nemirw@gmail.com)

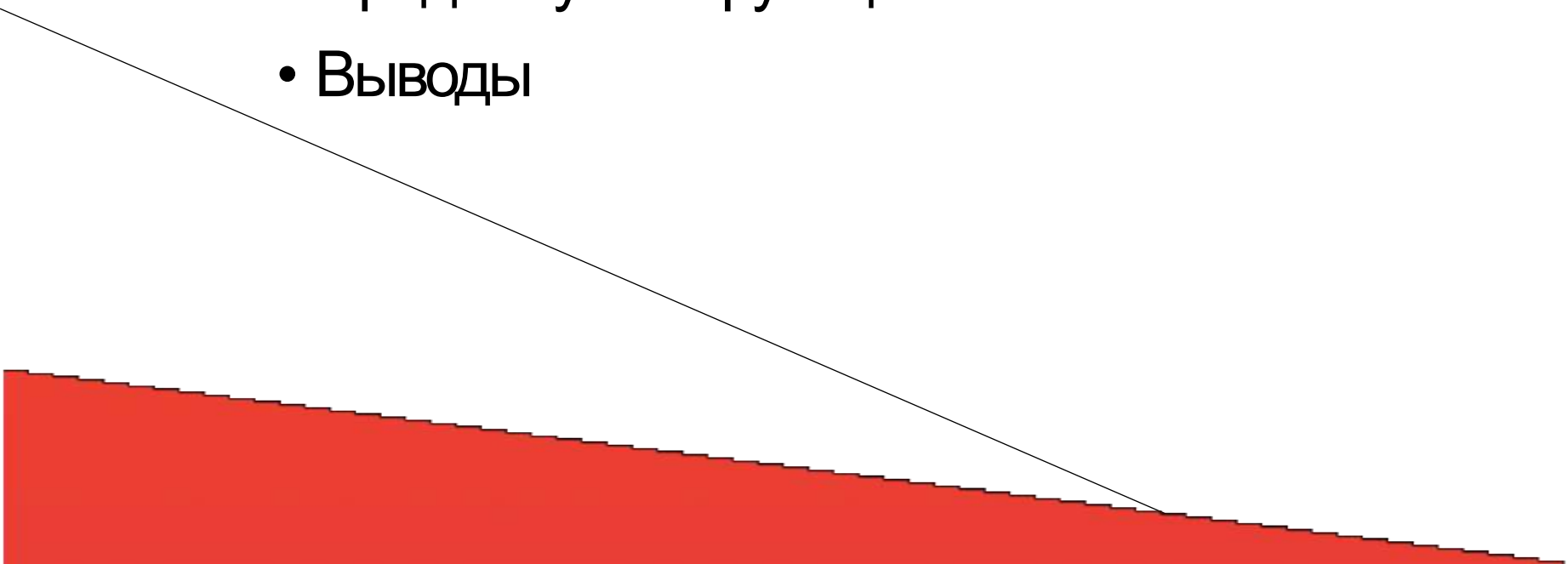
Елисей Сажин (sazhin.elisey@gmail.com)

Алена Утробина (kladiki@mail.ru)

Мария Чудновская (mlc999@list.ru)

Проект по КИЛИ

# Структура презентации

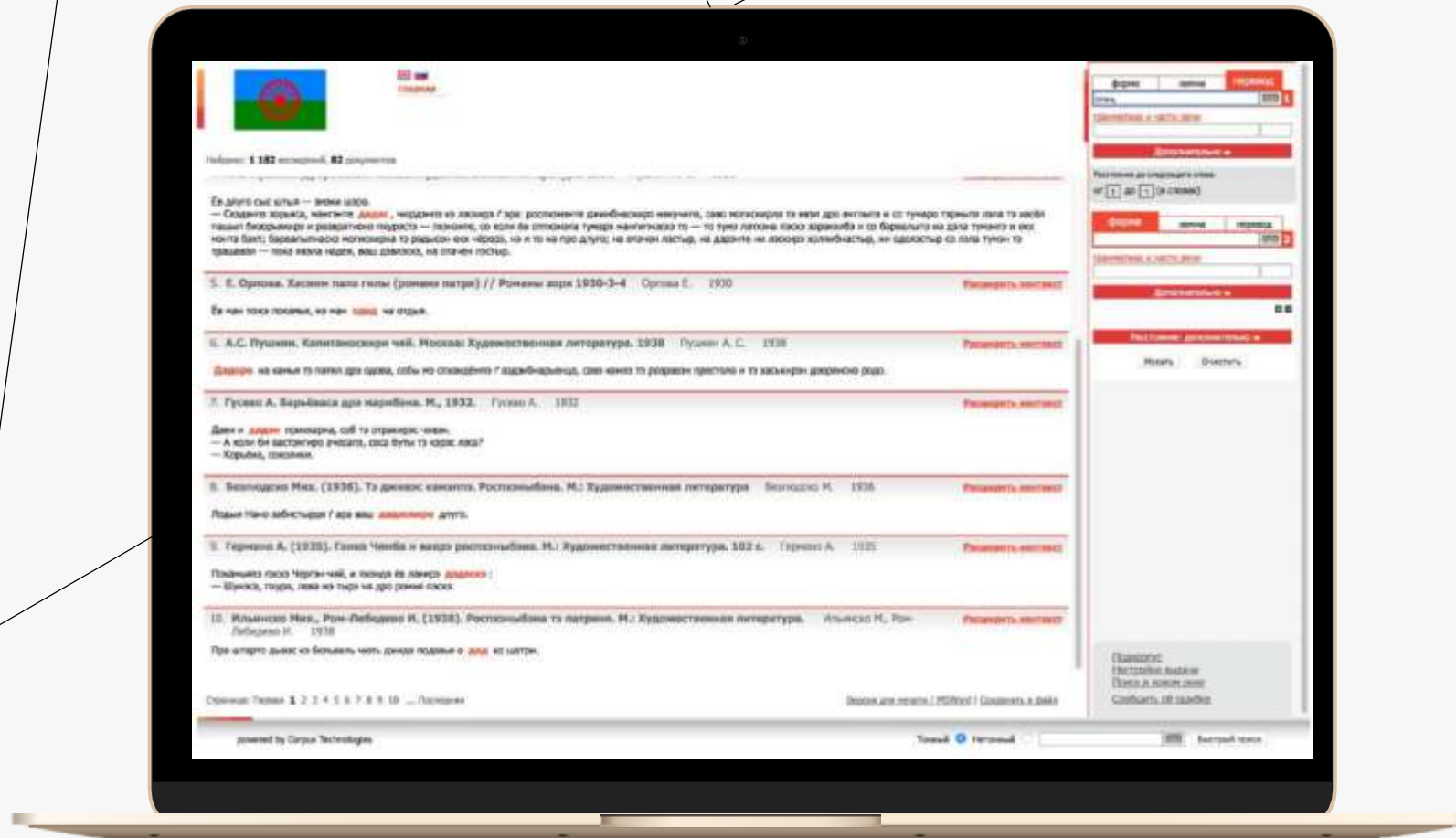
- Введение
  - Дизайн
  - Ресурс глазами новичка
  - Помощь пользователю
  - Продвинутый функционал
  - Выводы
- 
- A decorative graphic element consisting of a solid red triangle pointing upwards from the bottom left corner, and a thin black diagonal line extending from the top left towards the bottom right, intersecting the red triangle.

# Введение

В работе рассматривается Корпус цыганского языка (Russian Romani Corpus), который включает тексты, изданные в СССР в 1920-1930 гг. Корпус имеет несколько слоев разметки: метатекстовую (библиографическую) и грамматическую.

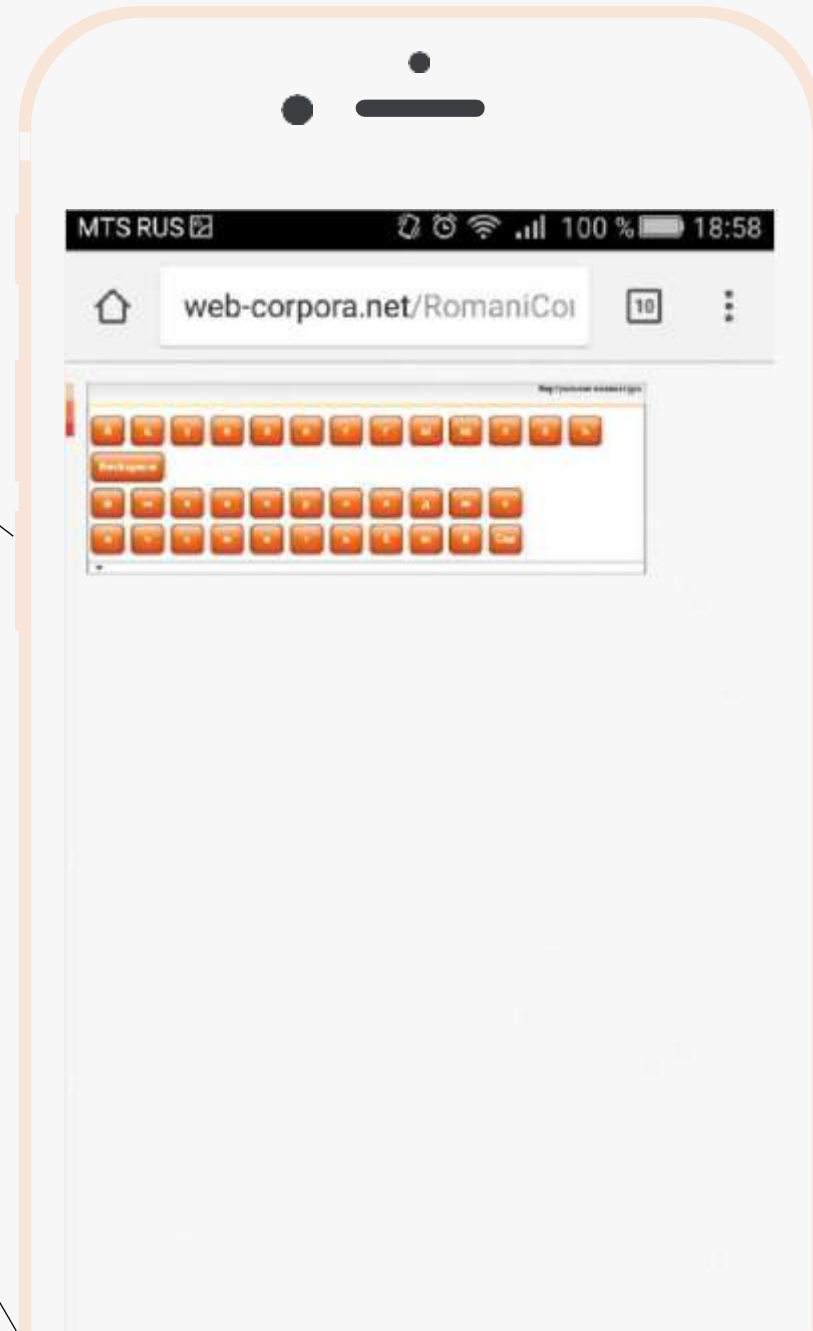
# Дизайн

- Количество вхождений и документов подсвечены жирным
- Информация об источнике выделена темно-серым цветом
- Искомое слово/лемма выделены оранжевым



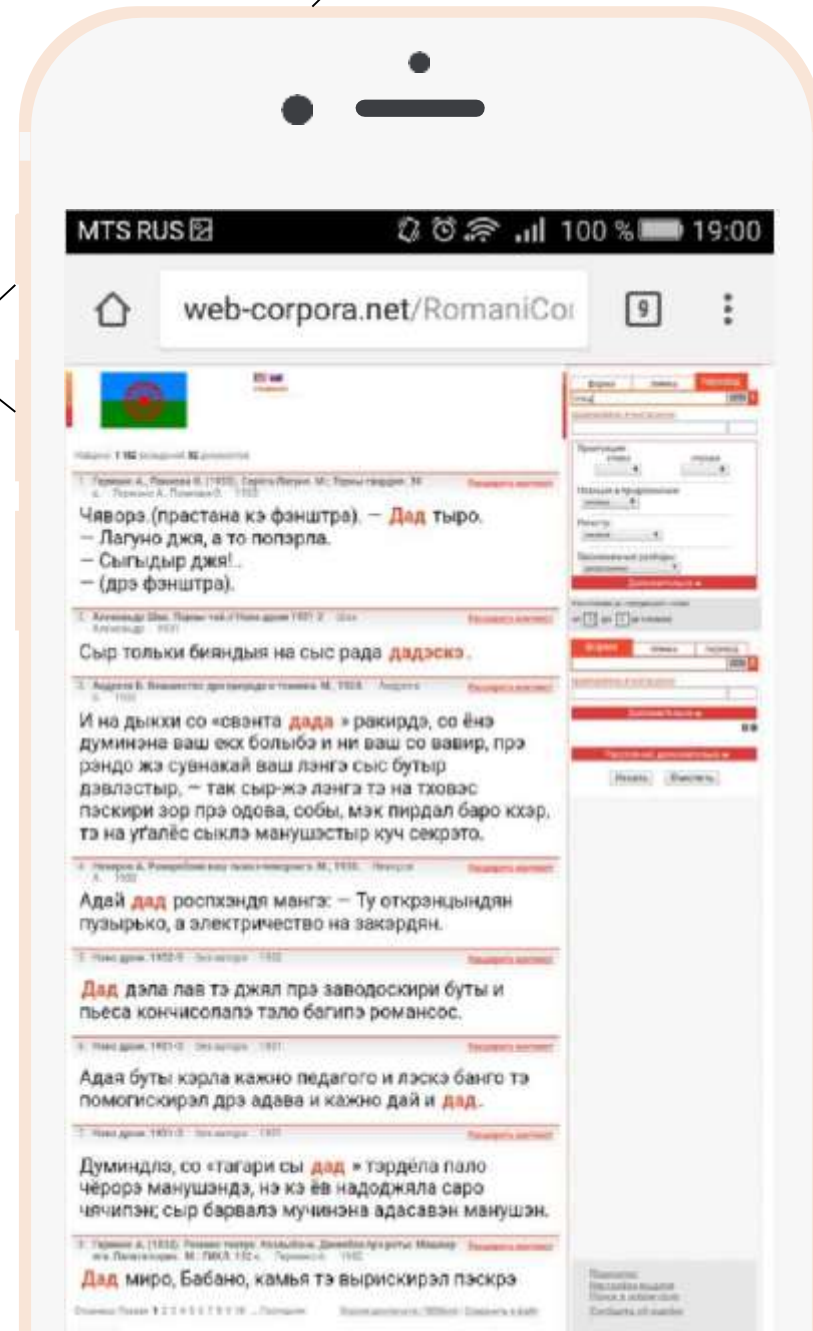
# Дизайн

- Неудобства при работе на мобильных устройствах
- Виртуальная клавиатура открывается в новой вкладке на весь экран



# Дизайн

- На небольшом экране страница выглядит чересчур сжатой
- Недоработанность дизайна VS практичность
- Советы: более лаконичное и сглаженное оформление, разделение на смысловые блоки





# 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.

# Ресурс глазами новичка

Вид  
главной  
страницы  
сайта:



## Romani corpus

This is the home page of the Russian Romani Corpus, which currently contains approximately 720,000 tokens. At the moment, the corpus only contains texts published in te USSR in the 1920s and 1930s. The corpus includes all original texts (both fiction and press), as well as a handful of translated texts (fiction, non-fiction and press). The corpus is still under development. Right now, the user can make lexical and grammatical queries in the corpus. In the future, we are going to increase the size of the corpus by including the rest of the texts published in the 1920s and 1930s, as well as by adding the texts collected during the fieldwork conducted by the authors of the corpus. We are also going to improve the quality of the morphological annotation, enlarge our grammatical dictionary and perform disambiguation.

The corpus is being developed by [K. Kozhanov \(Moscow\)](#), [S. Oskolskaya \(St. Petersburg\)](#), [M. Osion \(Moscow\)](#), [A. Tenser \(Helsinki\)](#), [T. Arkhangelskiy \(Moscow\)](#).

The corpus was annotated with the help of an automated morphological annotation tool *UniParser* developed by T. Arkhangelskiy. The search platform of the [Eastern Armenian National Corpus \(EANC\)](#) was used for this corpus. You can read about making search queries at [EANC help page](#).

The corpus has been developed with the support of the RFBR grant no. mol\_a 14-06-31038 "Development of a Russian Romani Corpus" headed by K. Kozhanov in 2014–2015 .

Wordform

Lexeme

Translation

1

Gram & Lexical Attributes

Advanced

Distance to the next token:

From 1 to 1 words

Wordform

Lexeme

Translation

2

Gram & Lexical Attributes

Advanced

Advanced Distance

Search

Clear

[Specify Subcorpus](#)  
[Display Options](#)  
[Search in New Window](#)  
[Error Report](#)

powered by Corpus Technologies

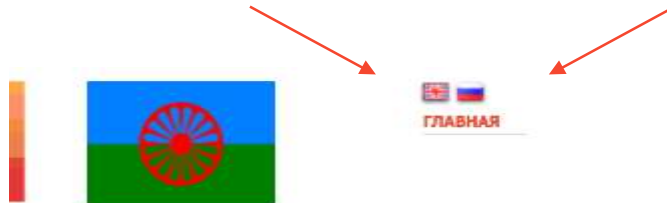
Exact ☒ Loose ☐

QuickSearch



# Ресурс глазами новичка

Переключение  
между англо- и  
русскоязычной  
версиями  
осуществляется  
путем нажатия на  
соответствующую  
кнопку-флажок  
вверху страницы:



Корпус цыганского языка

На этом сайте размещён Корпус языка русских цыган, насчитывающий 720 тыс. словоупотреблений. На данный момент он полностью состоит из текстов, изданных в СССР в 1920-х—1930-х годах. В корпус включены все оригинальные тексты (как художественные, так и публицистические), а также незначительное количество переводов (художественных, публицистических и научно-популярных текстов). Корпус находится в стадии разработки. Уже сейчас доступен поиск лексической и грамматической информации по имеющимся текстам. В дальнейшем планируется увеличить объем корпуса как за счёт остальных текстов, опубликованных в 1920-х—1930-х годах, так и за счёт полевых записей разработчиков корпуса. Кроме того, предполагается также улучшить морфологическую разметку, пополнить грамматический словарь и снять омонимию.

Работу над корпусом ведут [К. А. Кожанов \(Москва\)](#), [С. А. Оскольская \(Санкт-Петербург\)](#), [М. В. Ослон \(Москва\)](#), [А. Тенсер \(Хельсинки\)](#), [Т. А. Архангельский \(Москва\)](#).

Разметка корпуса совершена с помощью системы автоматического морфологического анализа UniParser, разработанной Т. А. Архангельским (Москва). Для корпуса была адаптирована поисковая система [Восточноармянского национального корпуса \(EANC\)](#), на сайте которого можно прочитать о том, [как пользоваться поисковым интерфейсом](#).

Корпус создан при финансовой поддержке гранта РФФИ мол\_а 14-06-31038 «Создание корпуса языка русских цыган» (рук. К. А. Кожанов, 2014—2015 гг.).

powered by Corpus Technologies

Точный ☒ Неточный ☐ Быстрый поиск

форма лемма перевод 1

грамматика и части речи

Дополнительно

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

форма лемма перевод 2

грамматика и части речи

Дополнительно

Расстояние: дополнительно

Искать Очистить

Подкорпус  
Настройки выдачи  
Поиск в новом окне  
Сообщить об ошибке

# Ресурс глазами новичка

- Настройки поиска
- Ссылка на создание подкорпуса
- Настройка выдачи результатов
- Интуитивно понятный интерфейс

The screenshot displays a web interface for a linguistic resource. At the top, there are three tabs: "форма" (form), "лемма" (lemma), and "перевод" (translation). Below these is a search bar with a red "1" icon. The interface includes sections for "грамматика и части речи" (grammar and parts of speech) and "Дополнительно" (Additional). A section titled "Расстояние до следующего слова:" (Distance to the next word:) shows "от 1 до 1 (в словах)" (from 1 to 1 (in words)). Below this is another set of tabs for "форма", "лемма", and "перевод", with a red "2" icon. The "Дополнительно" section includes a "Расстояние: дополнительно" (Distance: additionally) dropdown. At the bottom, there are buttons for "Искать" (Search) and "Очистить" (Clear). The footer contains links for "Подкорпус" (Subcorpus), "Настройки выдачи" (Output settings), "Поиск в новом окне" (Search in new window), and "Сообщить об ошибке" (Report error).

форма лемма перевод

грамматика и части речи

Дополнительно

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

форма лемма перевод

грамматика и части речи

Дополнительно

Расстояние: дополнительно

Искать Очистить

Подкорпус  
Настройки выдачи  
Поиск в новом окне  
Сообщить об ошибке

# Помощь пользователю

- Ссылка на инструкцию к поиску по Восточно-армянскому национальному корпусу
- Функции из ВАНК не всегда применимы к КЦЯ

Если мы попробуем действовать по инструкции пользователя ВАНК, то столкнемся с ошибкой:

Искомый элемент запроса не найден.  
Перевод **"\*give**

The screenshot displays the VANK (Vostочно-Армянский Национальный Корпус) search interface. At the top, there are three tabs: 'форма' (form), 'лемма' (lemma), and 'перевод' (translation), with 'перевод' being the active tab. Below the tabs is a search input field containing the text '\*give up'. To the right of the input field is a small keyboard icon and a red button with the number '1'. Below the input field, there is a section titled 'грамматика и части речи' (grammar and parts of speech) with two empty input fields. A red button labeled 'Дополнительно' (Additional) is located below this section. Below the red button, there is a grey box containing the text 'Расстояние до следующего слова: от 1 до 1 (в словах)' (Distance to the next word: from 1 to 1 (in words)). Below this, there are three tabs: 'форма' (form), 'лемма' (lemma), and 'перевод' (translation), with 'форма' being the active tab. Below the tabs is another search input field. To the right of this input field is a small keyboard icon and a red button with the number '2'. Below this input field, there is another section titled 'грамматика и части речи' (grammar and parts of speech) with two empty input fields. A red button labeled 'Дополнительно' (Additional) is located below this section. Below the red button, there is a red button labeled 'Расстояние: дополнительно' (Distance: additionally). At the bottom of the interface, there are two buttons: 'Искать' (Search) and 'Очистить' (Clear).

- Поиск по словоформе, по лемме, по грамматическим признакам, возможен поиск сочетаний из нескольких слов
- Возможность скачивания выборки
- Универсальность интерфейса
- Виртуальная клавиатура со специальными символами
- Вывод: корпус интуитивен для продвинутого пользователя

Помощь пользователю

# Продвинутый функционал

- Частотность использования существительных и местоимений в качестве подлежащего в конструкциях типа “SVO” в прозе и поэзии
- Подкорпус с прозой и поиск для подлежащего-существительного:

The screenshot displays a search interface for a corpus. At the top left, there is a logo with a red wheel on a green field. To its right is a button labeled "ГЛАВНАЯ". Below the logo, it says "Найдено: 81 вхождений, 37 документов" and "Размер подкорпуса: 94.23% от общего объема корпуса".

The search results are listed in a table with 8 entries. Each entry includes a number, author, title, year, and a link to "Расширить контекст".

№	Автор	Название	Год	Действие
1.	Германо А. (1932).	Романо театро. Кхэлыбэна. Джиибэн прэ роты. Машкир яга. Палага пэрво.	М.: ГИХЛ. 152 с. Германо А. 1932	<a href="#">Расширить контекст</a>
На дэ годла !...				
2.	Германо А. (1935).	Ганка Чямба и ваврэ роспхэныбэна. М.: Художественная литература.	102 с. Германо А. 1935	<a href="#">Расширить контекст</a>
Васта тринскирдэпэ холятыр .				
3.	Нэво дром. 1931-8	без автора	1931	<a href="#">Расширить контекст</a>
Пионеры гынэ кхэрэ .				
4.	Ляшко Н. Роспхэныбэн ваш случаё. М., 1935.	Ляшко Н.	1935	<a href="#">Расширить контекст</a>
Этажо кхэлдя ягэнца .				
5.	Нэво дром. 1932-2-3	без автора	1932	<a href="#">Расширить контекст</a>
Хась отлыджяла колхозостыр .				
6.	А.Г. Советско сэндэ // Романы зоря 1929-2	А. Г.	1929	<a href="#">Расширить контекст</a>
Табуно чёрдэ грэн .				
7.	Германо А. (1932).	Романо театро. Кхэлыбэна. Джиибэн прэ роты. Машкир яга. Палага пэрво. М.: ГИХЛ.	152 с. Германо 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 барьёла заводэнца ,				

At the bottom, there is a pagination bar: "Страница: Первая 1 2 3 4 5 6 7 8 9 ... Последняя". To the right of the pagination bar is a link "Версия для печати / MSWord" and a button "Сохранить в файл".



On the right side of the interface, there is a sidebar with advanced search options. It includes sections for "Пунктуация:" (left and right), "Позиция в предложении:" (beginning, middle, end), "Регистр:" (any), "Омонимичные разборы:" (allowed), and a "Дополнительно" button. There is also a section for "Расстояние до следующего слова:" (from 1 to 1 in words). The sidebar also has a "форма" section with tabs for "лемма" and "перевод", and a "грамматика и части речи" section.

At the bottom of the sidebar, there is a "Быстрый поиск" button.



# Продвинутый функционал

...и  
подлежащего-  
местоимения:

[ГЛАВНАЯ](#)

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

1. Светлово Л. (1938). Ром Хвасю. М.: Художественная литература. 136 с. Светлово 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>
Ёнэ сы чёра .	
3. Германо А. (1932). Романо театро. Кхэлыбэна. Джибэн прэ роты. Машкир яга. Палага пэрво. М.: ГИХЛ. 152 с. Германо А. 1932	<a href="#">Расширить контекст</a>
Сарэ дыкхэна э-рувэнца	
4. Л. Толстой. Трин рычя. Москва 1937 Ленинград Толстой Л. Н. 1937	<a href="#">Расширить контекст</a>
Вавир със дай .	
5. Светлово Л. (1938). Ром Хвасю. М.: Художественная литература. 136 с. Светлово Л. 1938	<a href="#">Расширить контекст</a>
Адава сы парамыся .	
6. Германо А. (1932). Романо театро. Кхэлыбэна. Джибэн прэ роты. Машкир яга. Палага пэрво. М.: ГИХЛ. 152 с. Германо А. 1932	<a href="#">Расширить контекст</a>
Саво мэ кулако ?	
7. Германо А. (1933). Лэс кхардэ рувэса и ваврэ роспхэныбэна. М.-Л.: Художественная литература. 119 с. Германо А. 1933	<a href="#">Расширить контекст</a>
Токо-со багандлэ гилы .	
8. Германо А. (1935). Ганка Чямба и ваврэ роспхэныбэна. М.: Художественная литература. 102 с. Германо А. 1935	<a href="#">Расширить контекст</a>
Вавир думиндя Ненила .	

Пунктуация: слева  справа

Позиция в предложении:

Регистр:

Омонимичные разборы:

[Дополнительно](#)

Расстояние до следующего слова: от  до  (в словах)

[форма](#) [лемма](#) [перевод](#)

[грамматика и части речи](#)

Пунктуация: слева  справа

Позиция в предложении:

Регистр:

Омонимичные разборы:

[Дополнительно](#)

Расстояние до следующего слова: от  до  (в словах)

[форма](#) [лемма](#) [перевод](#)

[грамматика и части речи](#)

Пунктуация: слева  справа

Позиция в предложении:

Страница: Первая [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) Последняя

[Версия для печати / MSWord](#) | [Сохранить в файл](#)

# Продвинутый функционал

Далее проделаем 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

Вывод: в предложениях типа “подлежащее-сказуемое- прямое дополнение” в роли подлежащего существительное выступает относительно чаще, чем местоимение.



# Продвинутый функционал

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 черного - нет
Есть перевод с английского	Выполнен не в полной мере	Рис. 2.1 Отсутствуют самые элементарные слова (mother, father)
В поиске есть возможность задавать омонимичные разборы	В наличии функция WITHOUT_GLOSS, значение неизвестно	
	Не снята омонимия	
Возможен поиск по времени издания произведений	Базу корпуса составляют тексты 1920-1930-х г.г.	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)	



# Приложение

[ГЛАВНАЯ](#)

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

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

Пало последня штардэша бэрша папуасэ удюхонэ, со машир **парнэ** ваврэ пхувитконэ манушэндэ, савэ явэна палэ морэстыр, дрэван набут мануша, савэ здэна прэ Маклаёстэ.

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

Окэ и ачём из пхуро, еюджини про **парно** савто.

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

Чячюно хулай дрэ полярно область — адава **парно** рыч, или, сыр лэс юкарна прэ северо ошкүё.

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

Лэстэ **парнэ** зоралэ данда, прэ упратуны ушт пиро ванглы выджяна калэ фэнды, о Илько чястэс кошэлапэ пхурэ ромэнца.

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

Адава със лакиро дад, пэскирэ полэ чямыенца и **парнэ** вэнсцнда и бакенбардэнца.

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

Ужэ **парно** о чён г' аздыяпэ учес и обчидя, сыр парно тхудэс, сари шатра, нэ Рахиль на пасия.

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

Прихордэ дивэл хэлэдэс Ягорийёс и приплэндэ тэ ухтэл лэскэ прэ **парнэ** грэстэ дро Дарыдаскиро таборо.  
— Ягорийё, угалёв, сосытэ Дарыдаскирэ рома нан на пригалёна? У бут времё ухтя пиро пхув Ягорийё, нэ на ногискирдэ тэ розродэл ромэн: ёв дро екх форо явэла, а ёнэ дро вавир уджяна.

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

А со ада юкэ лакирэ, сыр подысала прэ тутэ, адякэ и хасиян про **парно** свэта.

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 ☐

Быстрый поиск

форма  лемма  перевод

Белый  1

грамматика и части речи

Дополнительно

Расстояние до следующего слова:  
от  до  (в словах)

форма  лемма  перевод

грамматика и части речи

Дополнительно

Расстояние: дополнительно

Искать  Очистить

Подкорпус  
Настройки выдачи  
Поиск в новом окне  
Сообщить об ошибке

Рис. 1.1

# Приложение

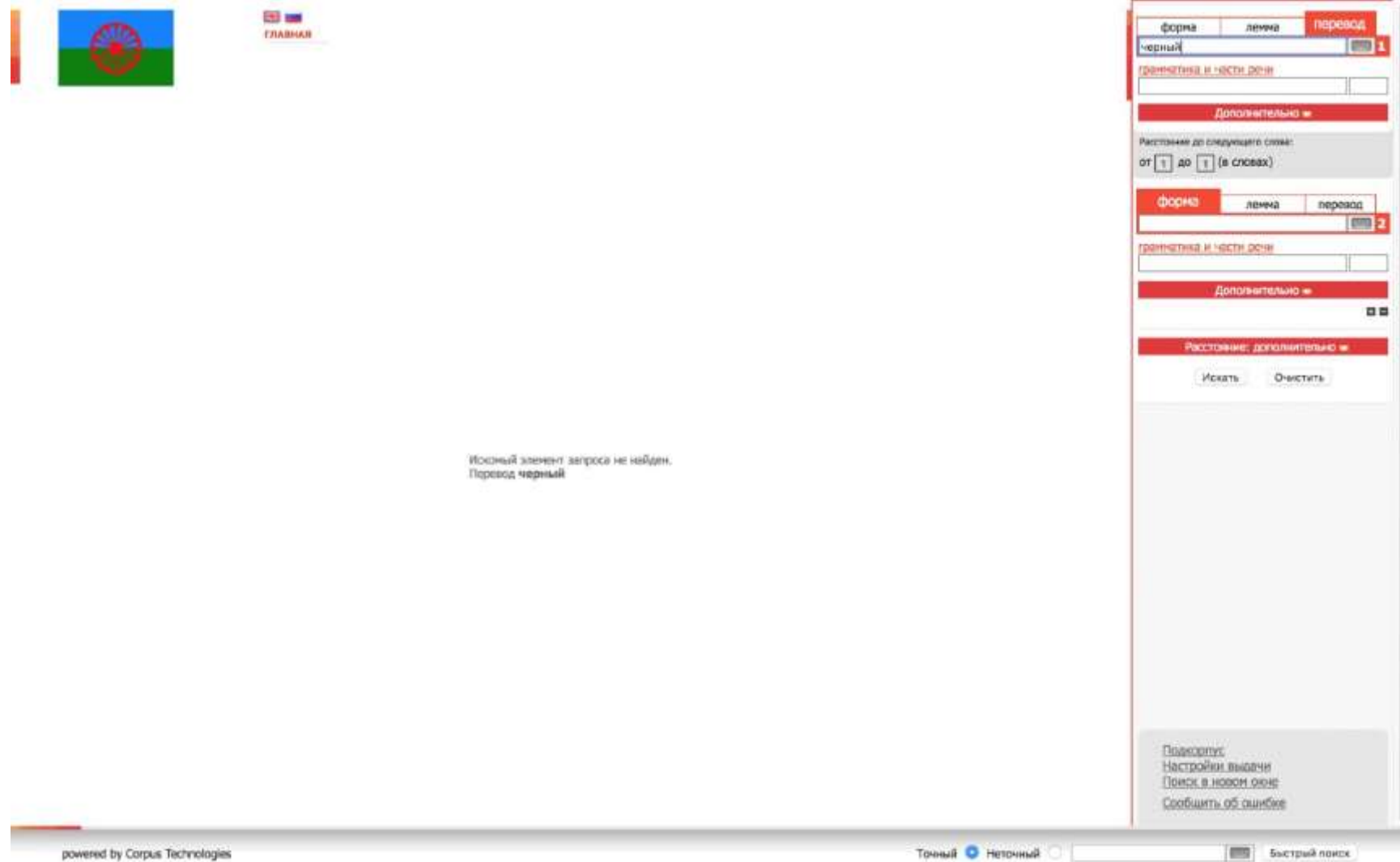


Рис. 1.2

# Приложение

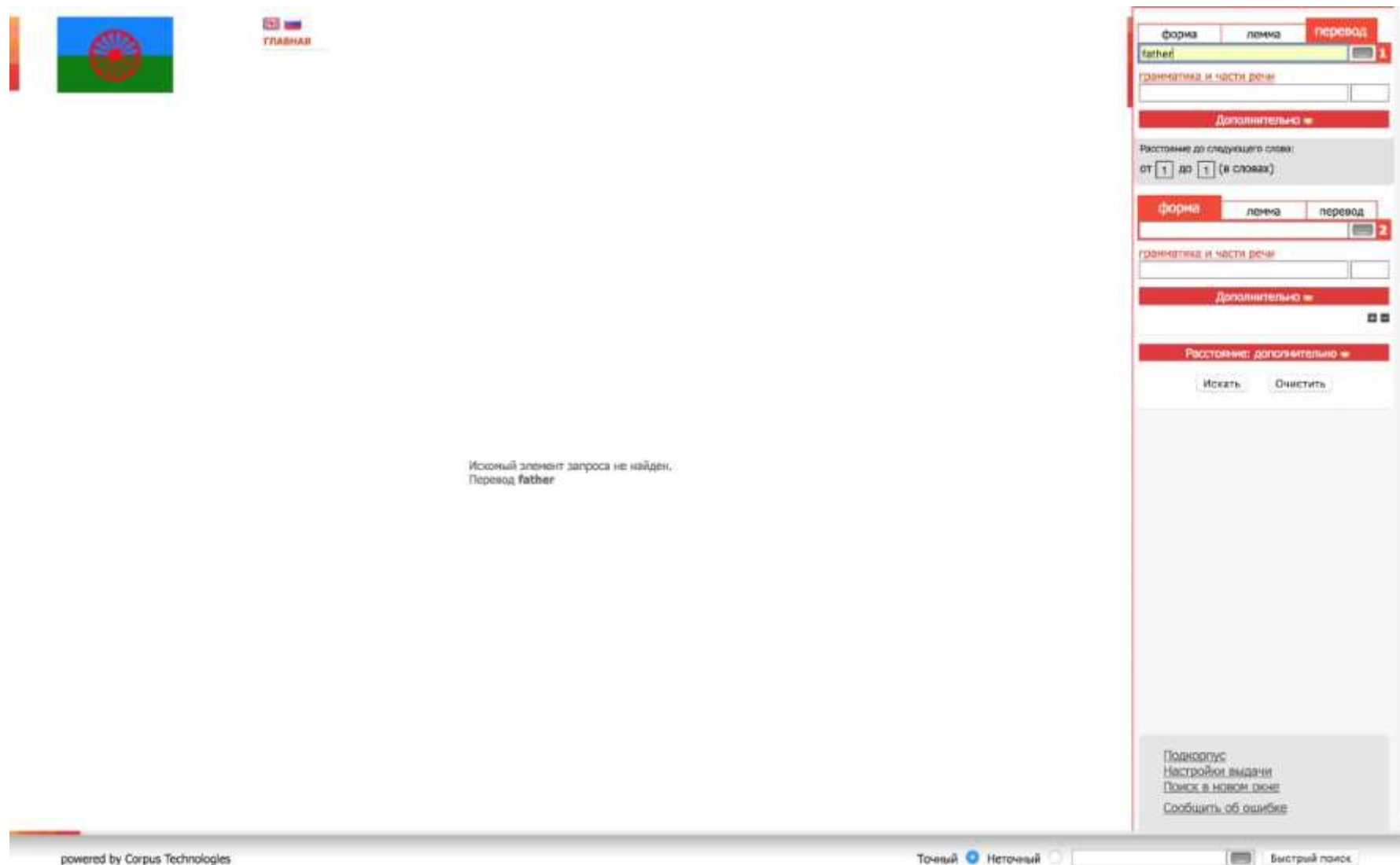


Рис. 2.1



# Выводы

- 
- Простота в применении
  - Недостаточно широкий функционал
  - Скучные возможности для работы в оффлайне
  - Примитивные запросы
  - Маленькая выборка текстов
  - Недочеты в дизайне

**Спасибо за  
внимание!**