Частотный словарь параллельного корпуса

Студенты:
София Землянская,
Аполлинария Карпова,
Александра Кузьмина,
Таисия Лукьянова,
Александра Нужненко,
Анастасия Сычева,
Анастасия Фирсова,
Елизавета Шемшурина

Руководитель:
Ольга Николаевна Ляшевская

Цель

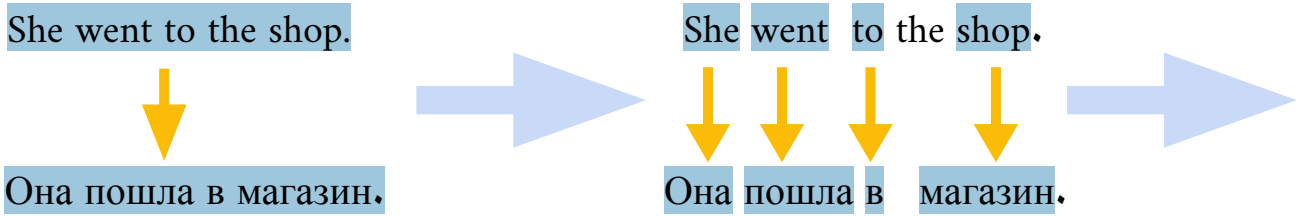
- Создать базу данных, основу для частотного словаря, в котором представлены **переводы** слова с одного языка на другой и **частотность** этих переводов. Информация о частотности основана на данных **параллельных корпусов.**

Для чего пригодится словарь

- **Лексическая типология**: теоретический анализ фреймов, подбор контекстов для опросов;
- **Психолингвистика**: при изучении билингвизма создаются эксперименты, в которых представлены одни и те же по смыслу предложения или слова на двух языках; словарь поможет подобрать наиболее подходящий перевод;
- **Преподавание иностранного языка**;
- **Машинный перевод**: на основе словаря можно создать датасет для оценки машинного перевода

Данные

- **Русско-английский** и **англо-русский корпуса**, в которых реализовано выравнивание по предложениям
- **Задача**: подготовить пословное выравнивание и частотные списки пар слов → пословное выравнивание обоих корпусов с помощью скрипта fast_align



eng	rus	num
she	она	n
go	пойти	m
to	в	k
shop	магазин	j

Обработка данных

- Предварительная обработка данных: преобразование файлов xml в txt формат, который поддерживает text_align: eng_text ||| rus_text или rus_text ||| eng_text;
- Апробация двух способов: сначала лемматизация, потом выравнивание и сначала выравнивание, потом лемматизация;
- Выбран вариант лемматизация - выравнивание, так как в нем было меньше мусора и больше адекватных соответствий (например, нет случаев, когда английский артикль приравнивается к русскому слову, так как артикли удаляются заранее)
- Добавление частей речи после слов;
- Английский – с помощью NLTK, русский – с помощью pymorphy2, теги nltk переделаны в pymorphy2 для однородности данных.

fast_align (C++, Cmake)

Формат входных данных:

doch jetzt ist der Held gefallen . ||| but now the hero has fallen .
neue Modelle werden erprobt . ||| new models are being tested .
doch fehlen uns neue Ressourcen . ||| but we lack new resources .

Alignment:

- source–target (left language–right language) alignments
- target–source (right language–left language) alignments
- atools: симметрия с использованием различных стандартных эвристик

fast_align (C++, Cmake)

Формат выходных данных:

- “Pharaon format” (i-j)

0-0 1-1 2-4 3-2 4-3 5-5 6-6
0-0 1-1 2-2 2-3 3-4 4-5
0-0 1-2 2-1 3-3 4-4 5-5

Создание списка слов:

8	0.0909	генрих	henry
1	0.0114	генрих	martin

Сложности выравнивания

- Неправильное соотнесение слов (например, в пару к самостоятельным словам подобраны служебные)
- Самый частотный перевод ≠ самый правильный

id	count	probability	english	russian
0	38	0.0219	bird	птица
1	19	0.0109	bird	птичий
2	71	0.0408	bird	и

Результат

Русско-английский корпус наречий - 11730 примеров;

Англо-русский корпус наречий- 22230 примеров;

Разметка результата: 0 - перевод неверный , 1 - перевод верный

index	count	probability	russian	english	result
25572	10	0.0075	трудно_ADVB	life_NOUN	0
25573	3	0.0022	трудно_ADVB	hard_ADVB	1
25574	3	0.0022	трудно_ADVB	began_VERB	0
25575	3	0.0022	трудно_ADVB	time_NOUN	0
25576	3	0.0022	трудно_ADVB	since_PREP	0
...					...

Русско-английский корпус: 1328 верных и 10402 неверных переводов

Сложности разметки

- Разные грамматические пометы у одних и тех же слов и слов, совпадающих по значению; соответствия разных частей речи.

index	count	probability	russian	english
86236	29	0.0122	тихо_ADVB	**quiet_ADJF**
86263	4	0.0017	тихо_ADVB	**quiet_NOUN**

- Случаи, в которых перевод — элемент соответствия;

index	count	probability	russian	english
103802	19	0.1	исподлобья_ADVB	looked_VERB
103806	3	0.0158	исподлобья_ADVB	beneath_PREP

- Обратные случаи: оригинал — элемент соответствия.

index	count	probability	russian	english
106201	3	0.0057	пристально_ADVB	stare_VERB

Контексты

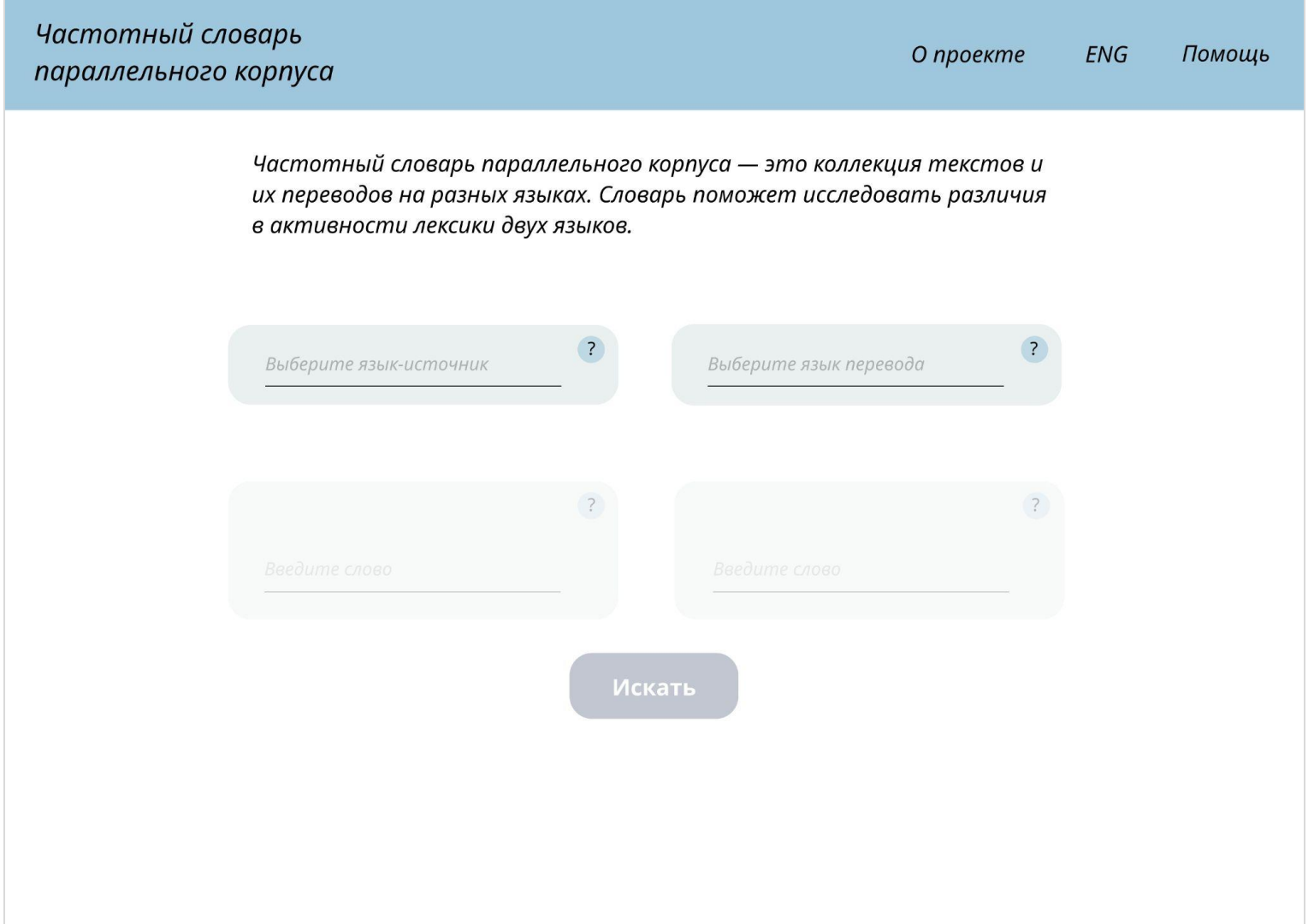
Как составлялась таблица:

- Для каждой пары из базы данных - поиск пар предложений, где есть оба слова

Но:

- Не всегда точные соответствия частотностям из базы
- Пока сделана только для русско-английского корпуса

Дизайн сайта



Частотный словарь параллельного корпуса — это коллекция текстов и их переводов на разных языках. Словарь поможет исследовать различия в активности лексики двух языков.

Русский ?

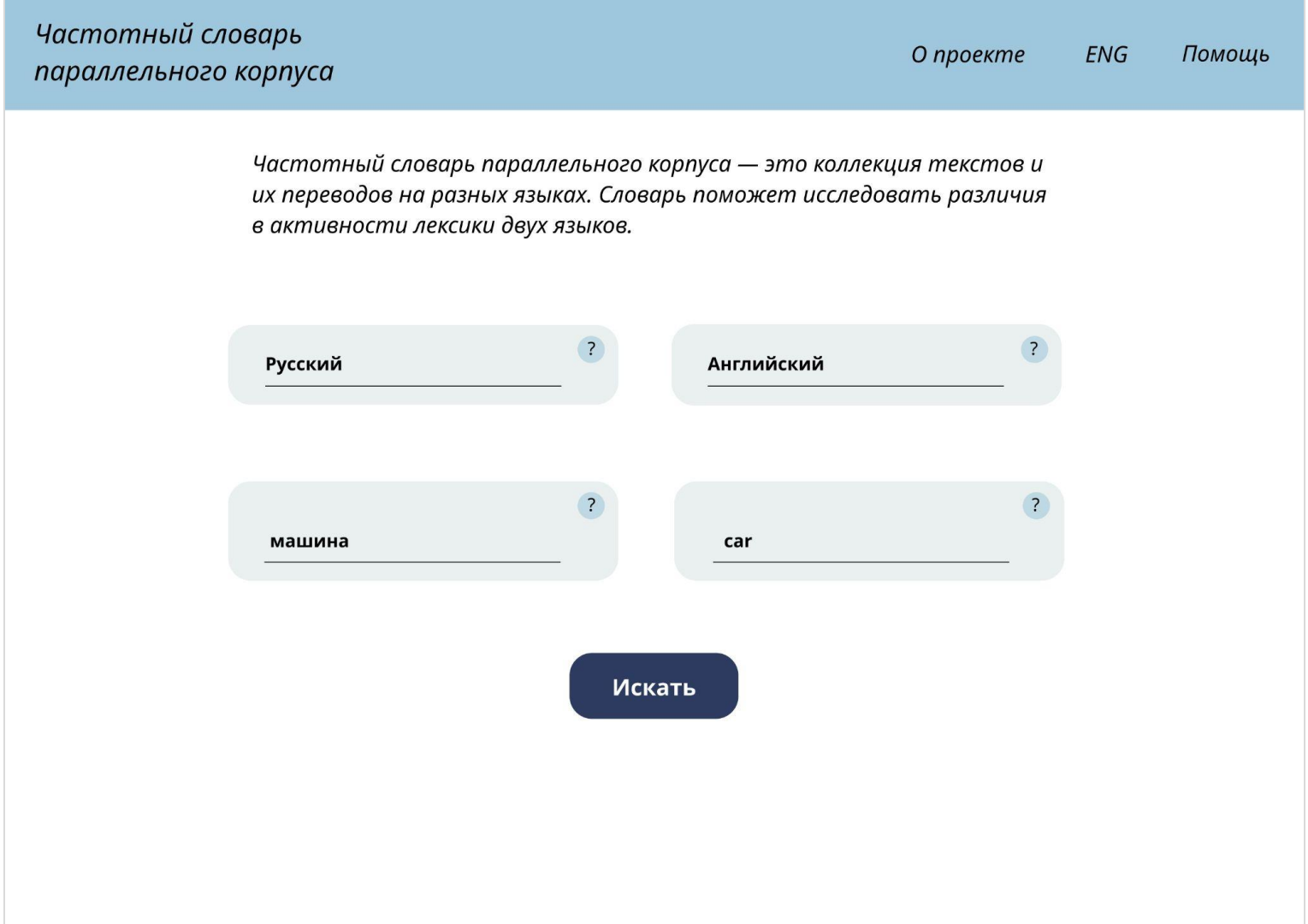
Английский ?

Введите слово ?

Введите слово ?

Искать

ПОИСК ПО ДВУМ СЛОВАМ



машина — car — 1056 hits

Скачать результат запроса

Машина *принадлежит Мелиссе Джой Блэк из Мичигана.*	*Uh,* ***car*** *belongs to a Melissa Joy Black from Michigan.*
Мы были женаты несколько лет, и у нас было двое детей и три ***машины****, но их он оставил себе.*	*We were married for several years and had two children and three* ***cars****, and he kept the cars.*
⋮	⋮
- И мы так думали, но ***машина*** *не застрахована.*	*"So we thought, but the* ***car's*** *not insured."*

<<< 25 из 1056 >>>

машина — car — 1056 hits

Скачать результат запроса

Машина принад
Мичигана.

Мы были женат
было двое дете
оставил себе.

- И мы так дума
застрахована.

ack from

and had two
pt the cars.

nsured."

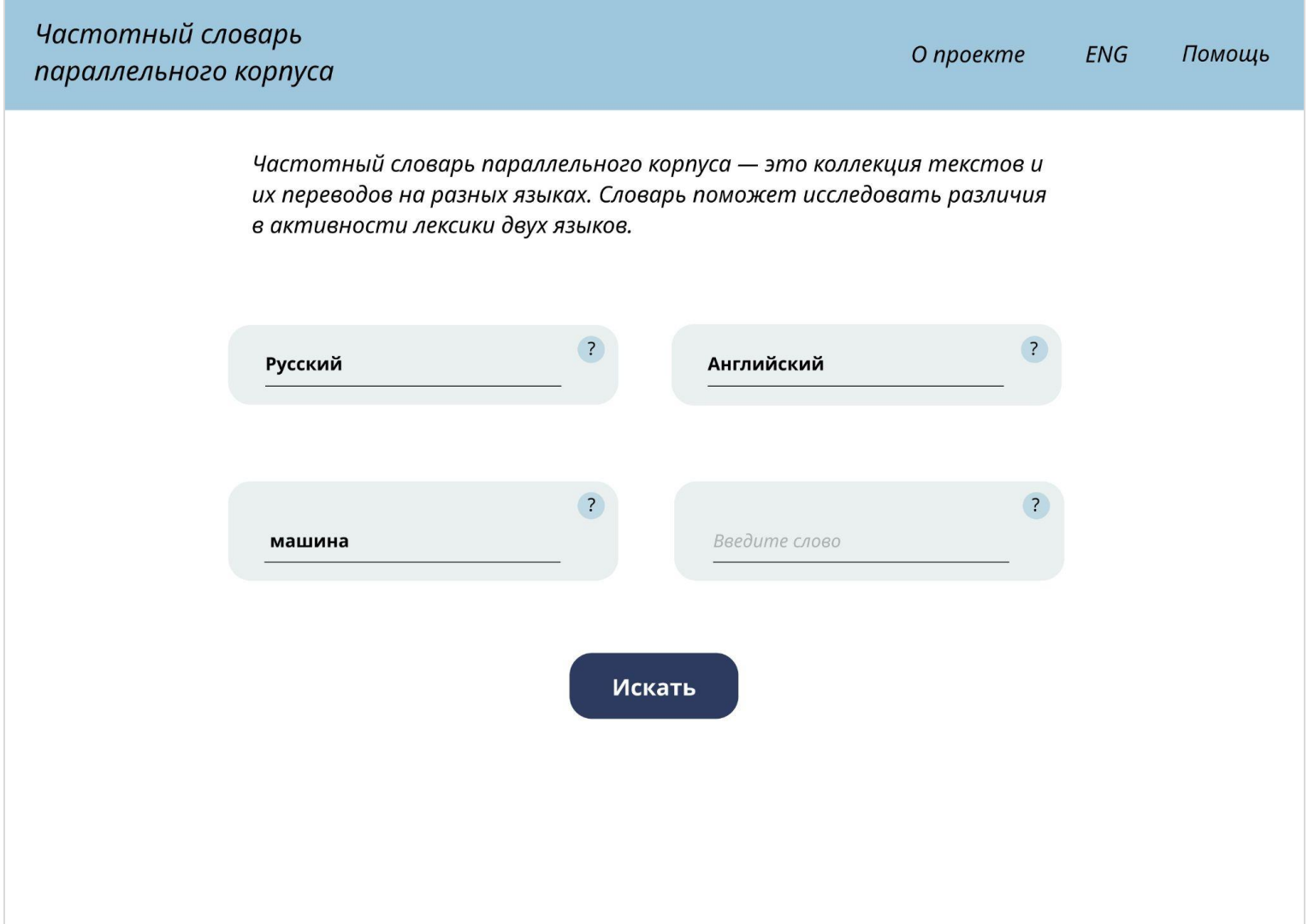
Сначала он размышлял о чем-то своем, затем медленно произнес:
- И мы так думали, но **машина** не застрахована.
Повисло напряженное молчание.

Автор: Василий Петров
Название: “Происшествие”

At first he was thinking about something of his own, then he said slowly:
"So we thought, but the **car's** not insured."
There was a tense silence.

Переводчик: Christopher R. Wood
Название: The incident

поиск по одному слову



Русский	Английский	Частота	
Машина *принадлежит Мелиссе Джой Блэк из Мичигана.*	*Uh,* ***car*** *belongs to a Melissa Joy Black from Michigan.*	? *1056 hits*	>>>
К полуночи ***машина*** *хорошо работала и полным ходом.*	*By midnight the* ***machinery*** *of his command was working thoroughly and efficiently.*	? *550 hits*	>>>
⋮	⋮		
— Бронированная ***машина*** *номер три. Мне необходимо срочно найти ее.*	*"Armored* ***truck*** *number three. I need to find it."*	? *35 hits*	>>>

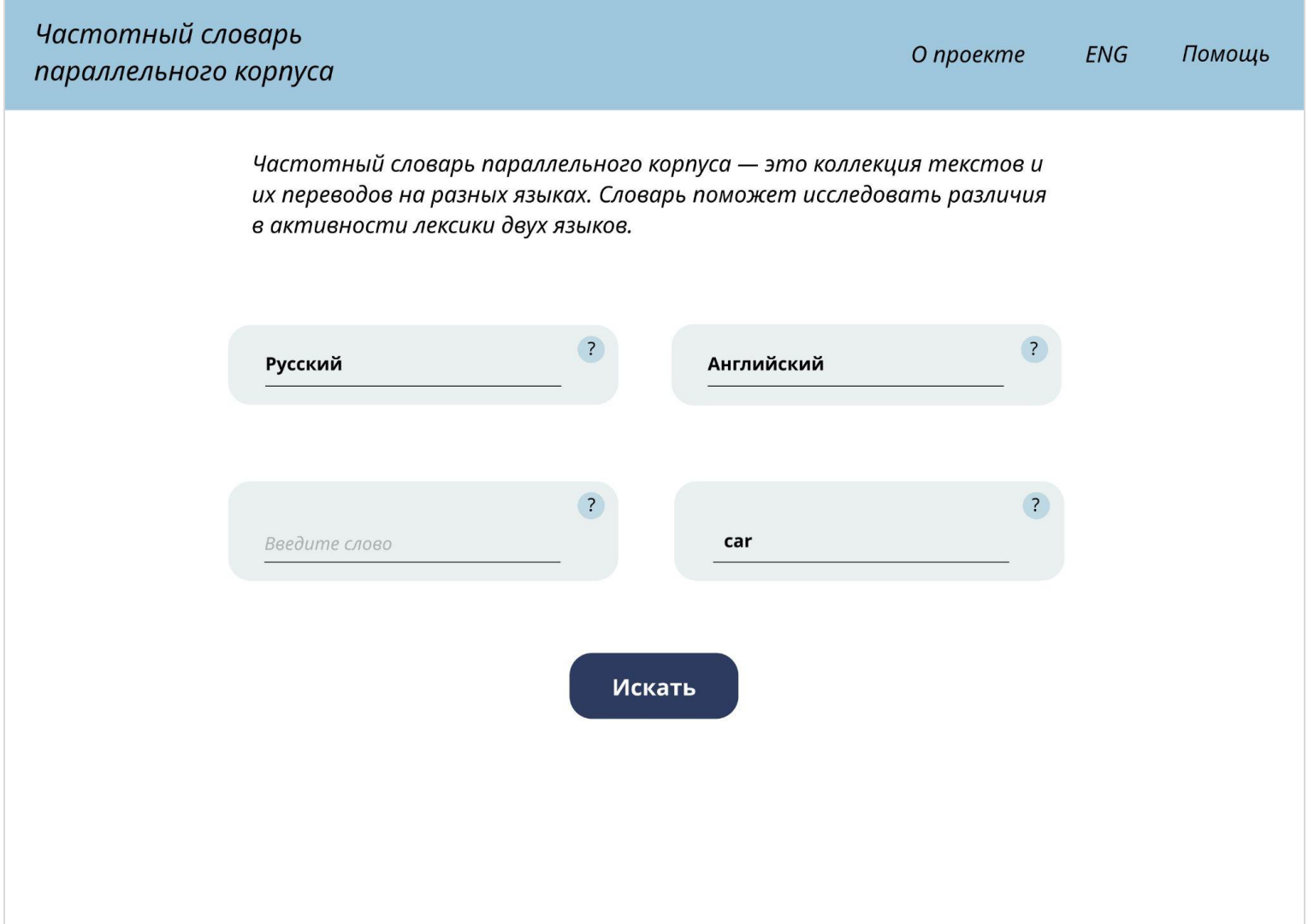
машина — car — 1056 hits

Скачать результат запроса

Машина *принадлежит Мелиссе Джой Блэк из Мичигана.*	*Uh,* ***car*** *belongs to a Melissa Joy Black from Michigan.*
Мы были женаты несколько лет, и у нас было двое детей и три ***машины****, но их он оставил себе.*	*We were married for several years and had two children and three* ***cars****, and he kept the cars.*
⋮	⋮
- И мы так думали, но ***машина*** *не застрахована.*	*"So we thought, but the* ***car's*** *not insured."*

<<< 25 из 1056 >>>

поиск по одному слову



Русский	Английский	Частота	
Она взяла напрокат **машину** и направилась к Лонг-Айленду. Она направлялась совершить ограбление.	She rented a **car** and headed for Long Island. She was on her way to commit a burglary. ?	1056 hits	>>>
А от **автомобиля** до космических полетов - всего лишь несколько десятилетий.	Yet only decades from the **car** into space. ?	550 hits	>>>
⋮	⋮		
Они отправились в путешествие в личном **вагоне** Джейми.	They made the journey in Jamie's private railway **car**. ?	35 hits	>>>

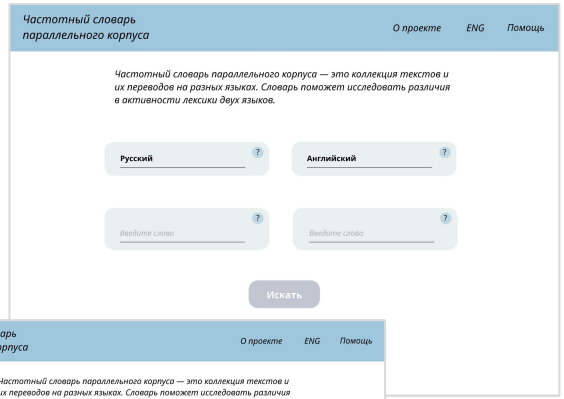
машина — car — 1056 hits

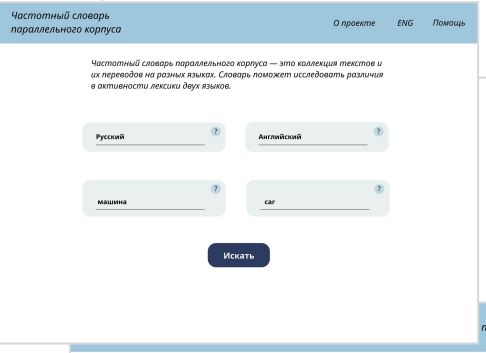
Скачать результат запроса

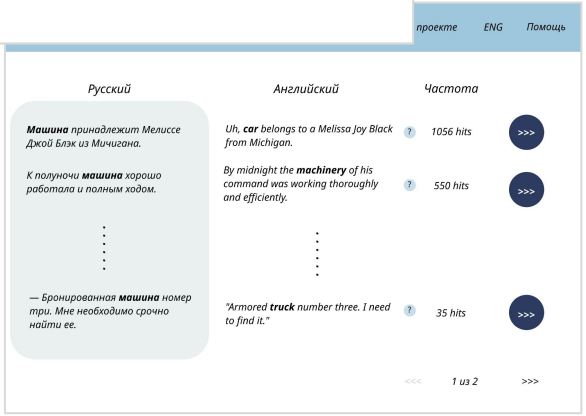
Машина *принадлежит Мелиссе Джой Блэк из Мичигана.*	*Uh,* ***car*** *belongs to a Melissa Joy Black from Michigan.*
Мы были женаты несколько лет, и у нас было двое детей и три ***машины****, но их он оставил себе.*	*We were married for several years and had two children and three* ***cars****, and he kept the cars.*
⋮	⋮
- И мы так думали, но ***машина*** *не застрахована.*	*"So we thought, but the* ***car's*** *not insured."*

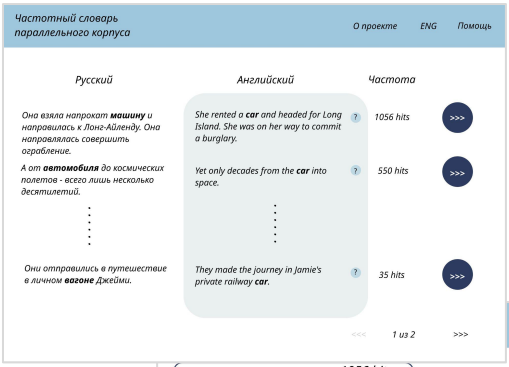
<<< 25 из 1056 >>>

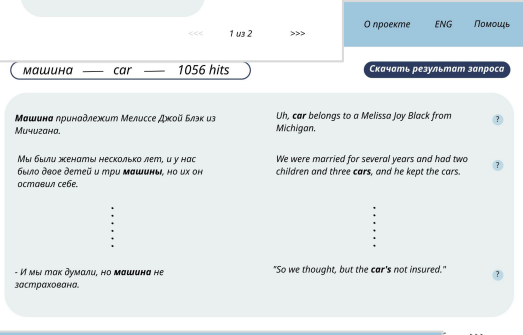
Макет сайта можно найти [здесь]

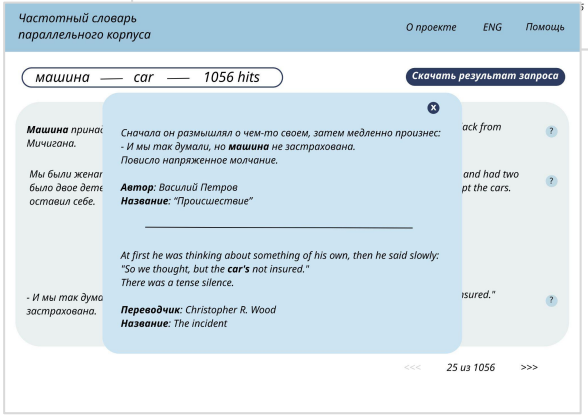












Перспективы проекта

- доделать сайт и выложить в общий доступ
- добавить другие пары языков:
  - русский-немецкий, русский-французский, русский-испанский и пр.
  - без русского: английский-немецкий, французский-испанский и пр.
- сохранять метаинформацию (автор, название произведения, год и др.)
- добавить тексты разных жанров
- снабдить некоторые из слов тэгами *разг.*, *вульг.*, *уст.* и пр.

Наша команда

- Аполлинария Карпова: автоматическое выравнивание;
- Анастасия Фирсова: автоматическое выравнивание;
- София Землянская: чистка и анализ данных;
- Александра Кузьмина: чистка и анализ данных;
- Таисия Лукьянова: чистка и анализ данных;
- Анастасия Сычева: чистка и анализ данных;
- Елизавета Шемшурина: чистка и анализ данных, подготовка контекстов для списков слово-перевод;
- Александра Нужненко: дизайн сайта.

Спасибо за внимание!