***Отчет об экспертизе корпусов Web-corpora***

В данной работе будет представлена оценка корпусов сайта Web-corpora на примере Бурятского, Калмыцкого и Албанского корпусов. При оценивании сайта учитывались следующие критерии: дизайн, удобство для пользователя, onboarding, помощь пользователю, функциональность.

**Дизайн**

Все 3 вышеуказанных корпуса имеют однотипный дизайн. В левом верхнем углу каждого корпуса находится небольшое изображение флага с надписью(Albanian National Corpus, Buryat Corpus, Kalmyk National Corpus соответственно). Под флагом находится текст, представляющий собой краткую информацию о корпусе, причем в Бурятском корпусе под общей информацией находится раздел “Как пользоваться корпусом?”, доступно описывающий, как пользоваться данным ресурсом. Справа находится непосредственно меню поиска с двумя полями ввода. Над обоими полями находится 3 кнопки: “форма”, “лемма” и “перевод”, под полями - дополнительное поле ввода и кнопка “дополнительно”. Также в нижней части страницы располагается “быстрый поиск” с собственным полем ввода.

Страница корпуса имеет белый фон, меню поиска оформлено в оранжевом цвете для Бурятского корпуса, в ярко-жёлтом цвете для Калмыцкого корпуса и в красном цвете для Албанского корпуса. Несмотря на яркость цветовой гаммы, цветовое оформление не является отвлекающим фактором и для глаза скорее приятно.

В целом, дизайн корпусов можно назвать минималистичным в связи с отсутствием большого количества декоративных деталей. Лаконичное оформление позволяет сконцентрироваться на основной задаче. Из минусов можно отметить неудобное расположение меню поиска на самом краю страницы, маленькие неприметные кнопки, наличие пустого пространства в верхней части страницы. Однако после некоторого количества произведённых запросов пользователь привыкает к этим особенностям, и дальнейшей работе они, как следствие, никак не мешают. Общее впечатление о работе с сайтом можно охарактеризовать как сдержанно-положительное.

Таким образом, сайт скорее удобен для работы, но к некоторым его особенностям нужно привыкать.

**Onboarding или ресурс глазами новичка**

Найти ресурс новичку или даже человеку, имеющему только базовое представление о корпусах, достаточно просто (2 вкладка в поиске браузера).

Все 3 корпуса имеют схожую схему работы(полностью соответствующую схеме работы Армянского корпуса, на основе которого и создавались корпуса). Первый поиск по корпусу не представляет особых сложностей, так как часть доступных функций интуитивно понятна. Однако требуется достаточно большое количество времени, чтобы начать использовать весь функционал сервиса. Сам поиск по корпусу может занимать значительный промежуток времени. К сугубо отрицательным моментам можно отнести отсутствие каких-либо инструкций и образцов запросов для новичков и статистических данных, таких как таблицы частотности.

В целом, корпуса сайта не отличаются удобством в работе, и легкими в освоении их не назовёшь.

**Помощь пользователю**

Из информации, представленной в корпусах, можно подчеркнуть, что все 3 вышеперечисленных корпуса сделаны на основе поисковой платформы “Восточноармянского национального корпуса”, ссылки на который содержатся во всех 3 корпусах. В самом Восточноармянском корпусе в главном меню содержится пункт “Как искать”, где очень подробно и понятно описан весь механизм поиска по корпусу. Из корпусов, про которые идёт речь в этом отчете, только в Бурятском корпусе есть раздел “Как пользоваться корпусом?”, в котором кратко описана схема работы с корпусом, что выгодно отличает его от Албанского и Калмыцкого корпусов.

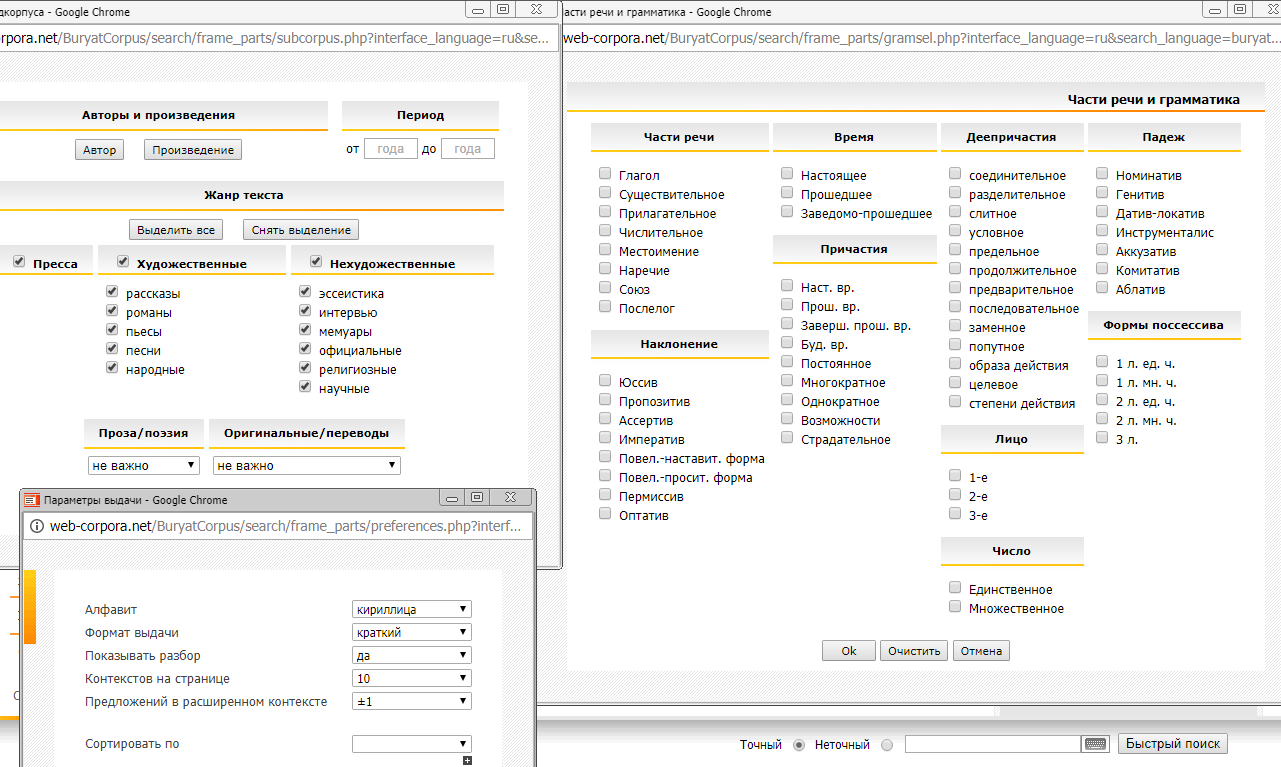
Таким образом, для того, чтобы найти раздел “помощь” необходимо перейти в Восточноармянский корпус по ссылке, что не очень удобно для начинающего пользователя, однако само описание(в Восточноармянском корпусе) составлено очень подробно и отвечает на все возможные вопросы.

**Функционал**

В данном пункте описан функционал Албанского, Бурятского и Калмыцкого корпусов в отдельности. Пункт содержит подпункты “Поиск по корпусам”, “Слои разметки”, “Текстовый состав и объем корпусов”, “Дополнительные возможности” и “Итоговая оценка возможностей корпусов”. В пункте “Поиск по корпусам” указана схема поиска по корпусам и ее краткая характеристика. В “Слоях разметки” подробнее описана разметка корпусов и приведена сравнительная таблица. В “Текстовом составе и объеме корпусов” кратко описан состав каждого корпуса. В “Дополнительных возможностях” находится информация о возможности скачивания данных из корпуса и оффлайн доступа. В “Итоговой оценке возможностей корпусов” присутствует описание различных функций корпусов и их оценка. В конце работы демонстрируются конкретные возможности поиска по корпусу и приведены примеры соответствующих поисковых запросов.

**Поиск по корпусам**

Как было отмечено ранее, поиск по корпусу основан на таковом в ВАНК. Параметры задаются через интерфейс поиска в правой части страницы. Ниже приведены окна со списками параметров, доступные по соответствующим ссылкам:

****

Также доступен поиск с использованием логических символов(| ~, для “или”, отрицания и “и” соответственно) и масок. Символы позволяют осуществлять поиск сразу нескольких слов(һүни , хамаг), слов по их окончаниям и началам(ха\*аг) и исключать результаты из выборки(~һүни). Следует отметить, что список этих функций и сам факт возможности их использования упомянуты только в туториалах ВАНК.

**Слои разметки**

|  |  |  |  |
| --- | --- | --- | --- |
| **Корпус\разметка** | **Бурятский** | **Калмыцкий** | **Албанский** |
| **Морфологическая** | есть(не всегда) | есть(не всегда) | есть(не всегда) |
| **Семантическая** | нет | нет | нет |
| **Омонимия** | нет | нет | нет |
| **Метаразметка** | есть(не всегда) | есть | есть |
| **Синтаксическая** | нет | нет | нет |
| **Перевод** | нет | есть(русский) | есть(английский) |

**Текстовый состав и объем корпусов**

*Албанский корпус*: содержит по большей части художественные тексты и публицистику; насчитывает около 20 млн. словоупотреблений.

*Бурятский корпус*: содержит по большей части художественные тексты, но с 2014 года появляются и образцы общественно-публицистического и учебно-научного стилей; объем - около 2 млн. 200 тыс. словоупотреблений

*Калмыцкий корпус*: в основном корпус состоит из художественных текстов; объем составляет около 800 тыс. словоупотреблений.

**Дополнительные возможности**

Ресурс не имеет оффлайн доступа, то есть доступен только при наличии сетевого подключения. Также отсутствует возможность полноценного скачивания данных из корпусов - скачанные материалы содержат только метаразметку по авторам. Эти моменты являются скорее отрицательными, так как дают работать только с неразмеченным материалом, сводя на нет многие преимущества корпуса.

**Итоговые возможности корпусов**

Все корпусы используют единую систему поиска, изначально разработанную для Восточноармянского Национального корпуса, из чего следуют особенности :

1. Рудиментарные функции и условия, оставшиеся от ВАНК: нерабочий перевод, поиск по устной речи без устной речи в корпусе, etc.
2. Одинаковый базовый набор настроек корпуса: знаки препинания, расстояние от-до, диапазоны поиска, пунктуация, названия простейших граммем, логические функции, etc.
3. Поиск по форме, лемме, переводу леммы, возможно задание граммемы, положения искомых слов друг относительно друга, поиск сочетаний.
4. Настройки выдачи включают контекст, разные форматы выдачи(глоссированный, KWIC, сортировка выдачи и скачивание).
5. Поиск с использованием функций тоже возможен, но такая возможность нигде не указана.

Таким образом:

|  |  |
| --- | --- |
| Можно | Нельзя |
| * найти базовую грамматическую информацию о языке * отслеживать последовательности слов и их свойства | * найти что-либо, связанное со значением слова * оффлайн обработать данные (не скачивается разметка) * найти сводные сведения (частотности употребления, временного распределения) |

Итого: разобранные и аналогичные им корпусы Web-corpora(корпусы иностранных языков и языков России) обладают довольно малым объемом(до 50 млн словоупотреблений - для сравнения, в НКРЯ - порядка 600 млн), слабой разметкой (ограниченная метаразметка и выполненная автоматически грамматическая разметка) и лишены механизмов отображения статистической и сводной информации, а потому подходят для изучения языков в отдельности только как источник необработанного материала. Преимущество же Web-corpora состоит в числе разных языков, что делает корпусы пригодными к использованию в качестве материала(пускай и поверхностного) для сравнительного анализа, а также уникальности представленных языков; кроме того, Web-corpora является удачной платформой для базовой реализации и дальнейшего развития корпусов малых языков.

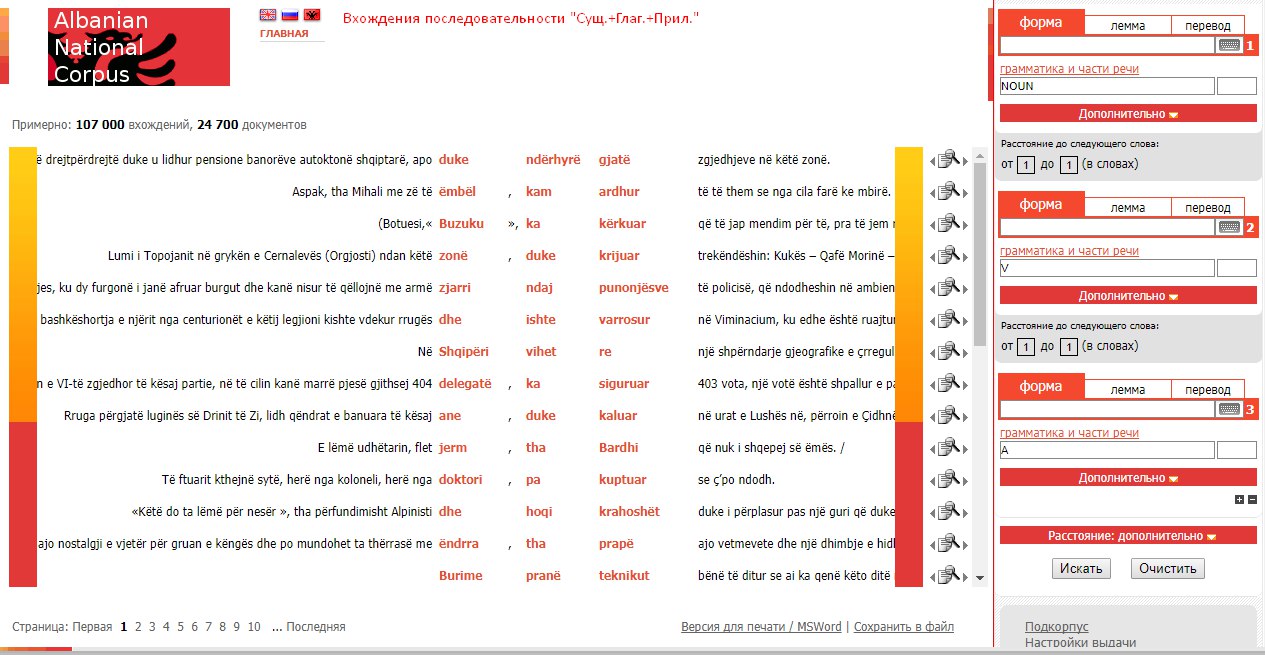
**Упоминания Web-corpora.org**

Данные корпусы в основном упоминаются как источник лингвистических сведений о малых языках и как примеры корпусов таких языков (<https://pdfs.semanticscholar.org/763a/2b42a64358cb9e06f994e17b5fb549e0fc44.pdf> )

Один из корпусов Web-corpora (Идиш) приводится в качестве примера успешной реализации корпуса “унаследованного языка” (heritage language): <https://www.researchgate.net/profile/Antonis_Botinis/publication/309727768_ExLing_2016/links/581fa5c408aea429b2990d56.pdf>

**Примеры поисковых запросов и выдачи**

*Запрос 1(Албанский корпус):*



*Запрос 2(Калмыцкий корпус):*



*Запрос 3(Бурятский корпус):*

**