Отчет об экспертизе usability

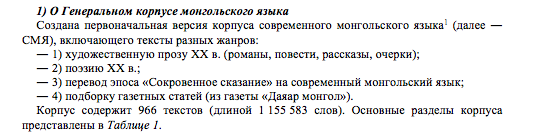
Монгольского национального корпуса  
(<http://web-corpora.net/MongolianCorpus/search/index.php?interface_language=ru>)

*Попова Рита, 1 группа*

*ritavpopova@gmail.com*

# Onboarding. Помощь пользователю

* Найти сам корпус не представляется проблемой - он представлен на сайте [web-corpora.net](http://web-corpora.net) и обнаруживается по первой ссылке в google по запросу “mongolian national corpus”.
* Для его использования не требуется регистрации, корпус доступен для онлайн-поиска.
* Макет страницы идентичен остальным корпусам [web-corpora.net](http://web-corpora.net), однако существенным отличием от большинства других является полное отсутствие сведений про сам корпус. Никакой дополнительной информации на его странице не содержится: она состоит исключительно из поисковика. Кажется, это наименее описанный из всех “корпусов других языков”, доступных на [web-corpora.net](http://web-corpora.net).
* При обращении ко внешним источникам можно довольно легко найти статью С. Крылова «Структурно-вероятностная модель монгольского языка (на базе Генерального корпуса современного монгольского языка)», в которой указана основная информация по корпусу, результаты работы с ним и т.д. Статья была написана в 2013 году и, кажется, описываемые в ней исследования, были проведены на материале, несколько отличающемся от содержащегося в корпусе на данный момент (например, его объем сейчас составляет 1160000 вхождений, как указано на web-corpora.net, а в статье С.Крылова называется число 1155583). Однако, приведу информацию из этой статьи, т.к. ничего другого найти, видимо, невозможно:



* Возвращаясь к странице корпуса, можем отметить, что ее интерфейс доступен на двух языках - русском и английском.
* Инструкции по использованию и подсказок не обнаруживается ни на одном из языков.
* Обратная связь с разработчиками возможна лишь через окно “сообщить об ошибке”, контактов или имен нигде не указано.
* Интерфейс предельно прост - главная страница совершенно пустая, на ней представлен лишь инструмент для поиска и логотип корпуса. Конструктор запросов организован очень компактно и интуитивно понятно: грамматические критерии для поиска словоформ можно прописать вручную или выбрать из таблицы, как и лексические категории. Хотя критериев, из которых можно выбрать, не так уж и много, значительно меньше, чем, например, для НКРЯ. Дополнительно можно выбрать место слова в предложении, регистр слова, отрегулировать омонимию. Выглядит несколько странно, как будто работа по оформлению не была завершена, однако функциональности это нисколько не мешает.
* Прямо на главной странице находится быстрый поиск
* Есть удобная кнопка, позволяющая задавать новый запрос в новом окне, это облегчает работу с корпусом.
* Более того, отсутствие подсказок и образцов запросов для начинающих пользователей не представляется большой проблемой - весь функционал находится на главной странице и предельно интуитивен в использовании; для того, чтобы разобраться в нем самостоятельно, не нужно много времени.

Так выглядит главная страница:



# Дизайн

* В оформлении страницы использованы всего 4 цвета, набор шрифтов минимален.
* Искомые слова в результатах запросов подсвечиваются желтым, что довольно удобно.
* Быстрый поиск находится прямо в футере и не выбивается из общего дизайна страницы.
* По непонятной причине большая часть результатов начинается с точки, а при расширении контекста видно, что между двумя предложениями вместо одной точки стоит две:

/Users/danya/Desktop/Screen Shot 2017-10-22 at 12.24.30.png

/Users/danya/Desktop/Screen Shot 2017-10-22 at 12.25.15.png

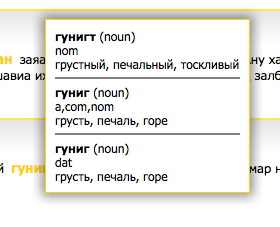
* Все кнопки имеют стандартный вид, такие используются для оформления большинства веб-страниц.
* Верстка страницы не выглядит особенно гармоничной из-за несоизмеримости ее элементов - огромная разница в размерах кнопок переключения языков и логотипа корпуса бросается в глаза столь же сильно, как незаполненность большей части пространства.
* В целом, дизайн страницы не вызывает дискомфорта, но стоит отметить его некоторую устарелость и незавершенность, которая присуща так же и функциональным возможностям корпуса.

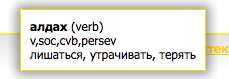
# Продвинутый функционал

* Помимо поиска лексем и словоформ, на странице представлен переводчик; переводы прописаны, однако, не для всех слов.  Более того, переводчик работает некорректно: “конь” переводится на монгольский (“морь”), а “лошадь” - нет, хотя в аннотации слова “морь” указаны оба перевода.
* Скачать в офлайн можно как все контексты, которые представлены на странице вместе, так и каждый из них по отдельности.
* Количество контекстов на странице так же регулируется.
* Поиск может быть осуществлен как по словоформе, так и по лексеме.  Количество слов в запросе может быть отрегулировано (точно можно задать до 10 слов, дальше не проверялось).
* Диапазон поиска может быть расширен с одного предложения до документа.
* Для всех результатов поиска можно расширить контекст и выбрать в нем количество предложений (до +3-х).
* Есть 7 вариантов сортировки результатов; сортировка при этом может осуществляться по нескольким параметрам одновременно:

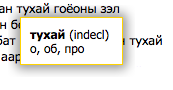


* При этом совершенно отсутствует метаразметка - не указаны автор, года создания и т.д. Судя по конструктору подкорпуса, в MNC содержатся различные жанры текстов и устная речь; переводы и оригинальные тексты; запись в новой и классической орфографии. Однако, это никаким образом не размечено на самих текстах.
* Грамматический поиск функционирует очень странным образом. Из всех частей речи он находит только глагол и существительное. То есть абсолютно все лексемы могут быть классифицированы либо как noun:

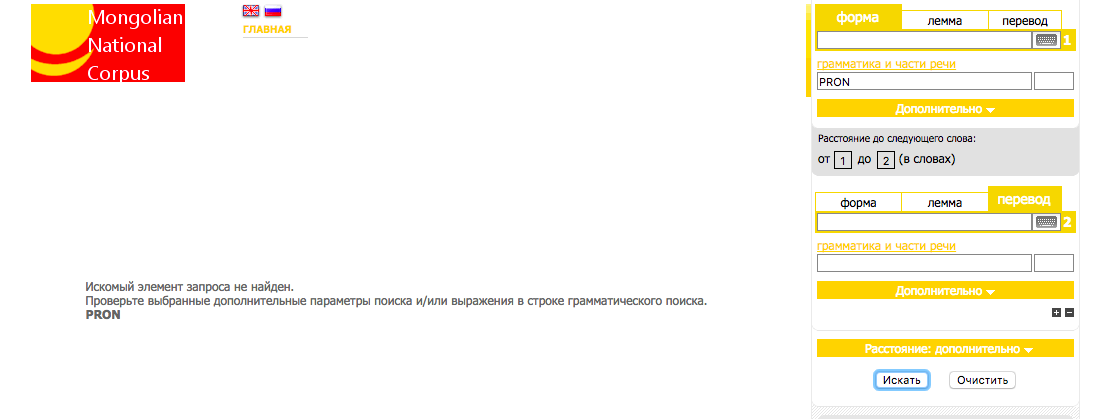
  
либо как verb:



либо часть речи вообще не определена:

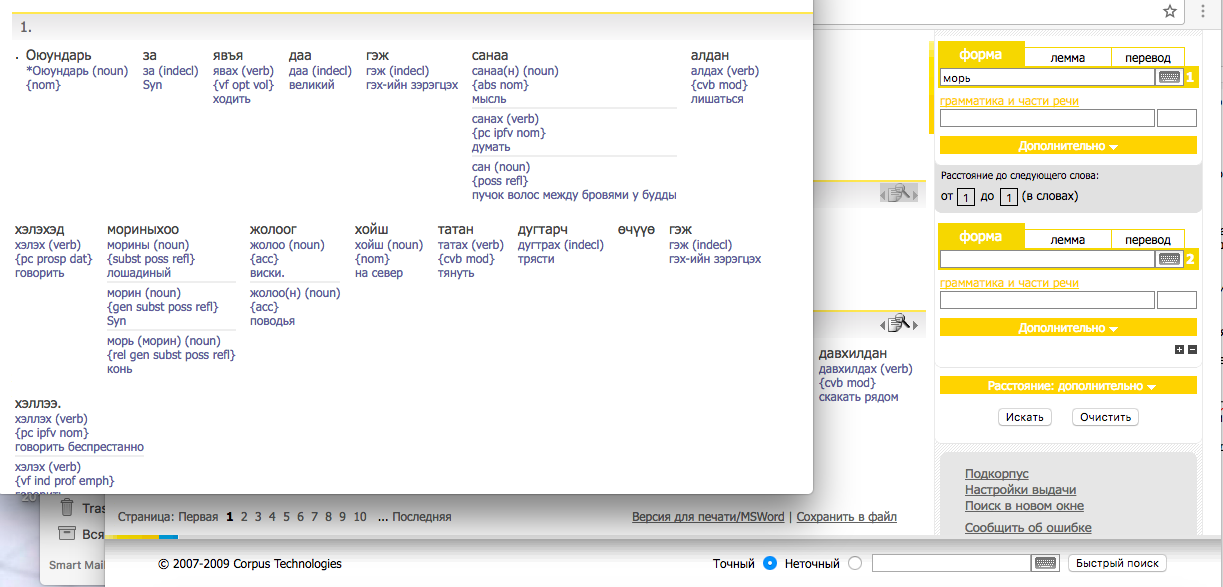


Числительное «первый», например, в одном из контекстов было размечено просто как «noun, nom». Как verb, при этом, во всех проанализированных контекстах были размечены правда только глаголы. По всей видимости причина кроется в неизвестном мне строе монгольского языка и все части речи кроме глагола являются именными и поэтому размечены как noun. В некоторых аннотациях помимо verb и noun под лексемой все же указывается настоящая часть речи (а, pron и др.), но никакой закономерности я выделить не смогла. Это, однако, не отменяет очевидных проблем в функционировании грамматического поиска, в котором можно выбрать любую из частей речи, которые вообще никак не указаны в разметках. Вот, к примеру, результат поиска любого местоимения:

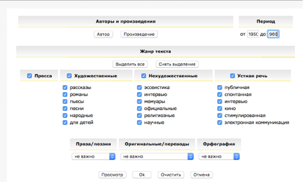


Соответственно, поисковые запросы из двух и более слов так же могут состоять только из verbs и nouns и потом анализироваться вручную.

* Аннотированы далеко не все словоформы: не размечены как редко встречающиеся словоформы («хожуухан», например, имеет 2 вхождения), так и относительно частые («зүүн» - 568 вхождений). Хотя никакой адекватной оценки частотности слов дать нельзя, потому что не известен сам объем корпуса.
* В разделе «формат выдачи» можно выбрать алфавит: армянский или транслитерацию. Почему алфавит армянский, не очень понятно, видимо это еще одна ошибка. Существуют корпуса еще для двух языков монгольской семьи – бурятского и калмыцкого. Там выбор алфавита осуществляется между кириллицей и транслитерацией.
* Формат выдачи, помимо полного, краткого и KWIC, может быть изменен на глоссированный – таким образом аннотация отображается прямо на странице:



* Значительной проблемой оборачивается любая попытка задать подкорпус - их было произведено 15, каждая из них по разным критериям, и каждый раз система выдавала ошибку. Например, я попыталась задать подкорпус, состоящий из любых текстов, написанных с 1950 по 1960 год:





# Вывод

Таким образом, не самый глубокий анализ функционала MNC буквально сразу же выявил его полную непригодность для работы. В нем есть возможность задать очень многие пользовательские настройки, что на первый взгляд весьма удобно, но большинство из этих функций просто не работают или вызывают вопросы, найти ответы на которые не представляется возможным. По всей видимости, работа над корпусом не была завершена или приостановлена, поэтому за вполне приемлемым внешним видом скрываются многие функциональные ошибки, из-за которых на материале корпуса становится невозможным провести какое-либо исследование.