АВТОМАТИЧЕСКОЕ ЗАПОЛНЕНИЕ БД ПОРТАЛА ТЮРКСКОЙ МОРФЕМЫ С ПОМОЩЬЮ ПРОГРАММНОЙ ОБРАБОТКИ ДВУЯЗЫЧНЫХ СЛОВАРЕЙ

Аюпов М.М.

Институт прикладной семиотики
Академии наук Республики Татарстан

19 октября 2020 г.

При создании больших многоязычных баз данных важная роль принадлежит заполнению этих баз данных, так как готовых решений не существует и подготовка нужных данных занимает много времени.

7 языков для портала тюркской морфемы:

– татарский,

– киргизский,

– узбекский,

– крымско-татарский,

– казахский,

– чувашский,

– башкирский.

Russian	Semant
бегун	1192

Tatar	Semant
йөгерешче	1192

Uzbek	Semant
yuguruvchi	1192

Tatar	Uzbek
йөгерешче	yuguruvchi

Tatar	Kirgiz	Uzbek	Cirim tatar	Rus	Kazah	Bashkir
абайлылык, ихтираз,	аярдык, кыраакылык,	tuyg'unlik	ачыкъкозьлик, ихтият,	осторожность	сезімталдық	
абайлы, сагаюлы, сак,	абайлагыч, кыраакы		ачыкъкозь, мукъайт,	осторожный	қырағы	һаҡ
йөгер, чап	жүгүр	chop, yugur	чап, ювур	бегать	шап	йүгерергә
бегемот, гиппопотам					сусиыр	
качкын	качкын, качуучу	qochkin	къачакъ, къачкъын	беглец	қашқын	
җәһәтлек, йөгереклек,		ildamlik, tezlik	сурьат, тезлик	быстрота	шапшаңдық	тиҙлек
качкан	качкы			беглый	қашқын	
алгасак, алгыр, алгысак,	амалдуу, бүкта,	chopkir, ildam,	сурьатлы, тез, чабик,	проворный	шапшаң	етеҙ, әлгәр, сос, уңған
бегония			бал-къаймакъ	бегония	бегония	
йөгерешче, йөгерүче		yugurdak,		бегун	жүгіріш	
афәт, бәла, бәла-каза,	алат, апат,	balo, baxtsizlik,	бахытсызлыкъ, беля,	беда	сорлылық	бәлә
болама, буталчыклык,	бүдүңчаң, иретсиздик,	sistemasizlik	къарышыкълыкъ,	беспорядок	жүйесіздік	бола
банкротлан, бөл,	жакырдан, жардылан,		сюмеле, факъырлаш,	беднеть	кедейлен	бөлөргә, хәлһеҙләнергә
йолкышлык, мәхрүмлек,	бакырдык, бакырлык,	faqirlik, muhtojlik	ёкъсуллыкъ, фукъарелик	бедность	мұқтаждық	
фәкыйрь-фөкарә, ярлы-				беднота	кедейлер	
барлыксыз, малсыз,	жакыр, жарды, кедей,	faqir	джарлы, ёкъсул, зюгюрт,	бедный	тіленшек	хәлһеҙ, ярлы
бәхетсез, кайгы-	бактысыз, бакытсыз,		бедбахт, гъарип,	невезучий	сорлы	
бахыр, бичара,	бечара		байгъуш, бичаре	бедняга	бишара	меҫкен
гидай	бакыр, кембагал			бедняк		фәкир
бот, сан	бут, сан	son		бедро	сан	бот

Язык	Количество слов
татарский	33066
казахский	18757
башкирский	2977
крымско-татарский	7065
киргизский	9686
узбекский	5433

Наиболее полные двуязычные словари в основном были созданы и изданы в прошлом веке. В связи с этим у них отсутствует электронная версия.

1. Найти отсканированную версию нужного словаря или, если нет версии с хорошим качеством, отсканировать.

2. Двуязычный словарь распознается с помощью специальных программ распознавания.

АККУРА`ТНО нареч. ұқыпты,мұқият,жинақы; ~ отвечать на письма хатқа мұқият жауап беру; ~ переписатьмұқият көшіру; 2. разг. абайлап,байқап,жайлап,білдірмей,ақырын; ~неситеабайлап апарындар; ~узнайжайлап біл; 3. разг. жүйелі,үнемі,үздіксіз; ~ навещать больногосырқатқа үнемі барып тұру.

3. С помощью программной обработки удаляется ненужная в дальнейшем информация.

АККУРАТНО ұқыпты,мұқият,жинақы 2. абайлап,байқап,жайлап,білдірмей,ақырын 3. жүйелі,үнемі,үздіксіз.

4. Обработанный словарь готовится для загрузки в базу данных.

АККУРАТНО	ұқыпты	1
АККУРАТНО	мұқият	1
АККУРАТНО	жинақы	1
АККУРАТНО	абайлап	2
АККУРАТНО	байқап	2
АККУРАТНО	жайлап	2
АККУРАТНО	білдірмей	2
АККУРАТНО	ақырын	2
АККУРАТНО	жүйелі	3
АККУРАТНО	үнемі	3
АККУРАТНО	үздіксіз	3

Казахский язык:

Экспериментальный словарь содержал **18757** строк.

Отсканированный pdf файл русско-казахского словаря содержал около 56000 словарных статей.

Готовый к загрузке в БД словник содержит более **117000** строк.

Возникают орфографические ошибки, если встречаются неуверенно распознанные слова и слова отсутствующие в словаре программы распознавания.

Отсутствие пробелов – распространенный случай при работе с распознанным текстом.

АБСОРБИ`РОВАТЬСЯсов.,несов. сорылу,сіңу,жұтылу.

АБСО`РБЦИЯж. физ., хим. абсорбция; сіңірілу,жұтылу, сорылу.

АБСТРАГИ`РОВАНИЕ с.абстракциялаушылық.

АБСТРАГИ`РОВАТЬсов., несов. что абстракциялау,дерексіздендіру,жалпылау.

АБСТРАГИ`РОВАТЬСЯ сов.,несов. абстракциялану,дерексіздену.

АБСТРА`КТНОСТЬж. дерексіздік,абстрактылық.

АБСТРА`КТН||ый,-ая,-ое абстрактылы,дерексіз; ~ые понятиядерексіз ұғымдар; ~ое тождествоабстрактылы тепе-теңдік.

Во время обработки словарей знак переноса строки обычно означает начало новой статьи. Но так же встречаются лишние знаки переноса строки.

ПЕРЕОБУЧА`ТЬ несов. см. переобучить.¶

ПЕРЕОБУЧА`ТЬСЯ несов. 1. см. переобучиться;¶

2 страд. отпереобучать.¶

ПЕРЕОБУЧЕ`НИЕ с. см. переобучить, переобучиться.-¶

ПЕРЕОБУЧИ`ТЬ сов. кого-что қайтаоқыту, жаңаданүйретіпшығару.¶

В словарях часто встречаются толкования-отсылки, которые требуют дополнительной обработки.

ПЕРЕРАБО`ТА||ТЬСЯсов.·1.·бойғасіну,·денегетаралу;·пища~ласьтамақбойғасінді;·2.·см.переработать5.¶

ПЕРЕРАБО`ТК||Аж.1.·см.·переработать1–4;·2.·разг.артықістелгенуақыт;·уплатитьза~уартықістелгенуақытүшінтөлем;·3.·(то,·чтопереработано)істепшығарылғанөнім,·ұқсатылғанбұйым.¶

ПЕРЕРАБО`ТОЧН||ЫЙ·~ая,·~оеұқсататын,·өңдейтін,·қайтажасайтын;·~ыйпунктөңдеупункті.¶

ПЕРЕРАСПРЕДЕЛЕ`НИЕс.см.перераспределить.¶

Выводы:

– получилось сэкономить время и трудозатраты при подготовке данных для загрузки в базу данных портала тюркской морфемы;

– хорошо бы разработать универсальное приложение для обработки двуязычных словарей с возможностью настройки работы программы через интерфейс.

Спасибо за внимание!

Игътибарыгыз өчен рәхмәт!