

Лингвистический корпус крымскотатарского языка

Radovan Garabík

Институт языкознания им. Л. Штура
Братислава, Словакия

Ленара Шакировна Кубединова
Таврический национальный университет им. В. И.
Вернадского
Симферополь, Украина

Крымскотатарский язык

- принадлежит к тюркским языкам
- кыпчакско-половецкая подгруппа кыпчакских языков
- средний диалект (на основе которого возник литературный язык) занимает положение между кыпчакскими и огузскими языками

Орфография

- до 1928 г. – арабский алфавит
- 1928 – 1939 – латинский алфавит
- 1939 – кириллица
- с 1997 г. – постепенный переход на новый латинский алфавит
- издаваемые тексты преимущественно на кириллице

Крымскотатарская кириллица

А а Б б В в Г г Гъ гъ Д д Е е Ё ё
Ж ж З з И и Й й К к Къ къ Л л М м
Н н Нъ нъ О о П п Р р С с Т т У у
Ф ф Х х Ц ц Ч ч Дж дж Ш ш Щ щ Ъ ъ
Ы ы Ь ь Э э Ю ю Я я

Крымскотатарская латиница

A a Â â B b C c Ç ç D d E e F f G g
Ğ ğ H h I ı İ i J j K k L l M m
N n Ñ ñ O o Ö ö P p Q q R r S s
Ş ş T t U u Ü ü V v Y y Z z

Корпус

- тексты в кириллице
- система клиент/сервер – корпусный менеджер Manatee/Bonito
- самостоятельный пользовательский веб интерфейс
- поиск регулярных выражений
- корпус доступен на интернете:
<http://korpus.juls.savba.sk/QIRIM>

Регулярные выражения

- *бала* найдёт все соответствия с единичным токеном *бала (ребёнок)*
- *бала.** найдёт все токены начинающиеся с *бала-*, т.е. *бала, балалар, балаларны, балачыкъларнен* и т.д.
- *кучюк бала* найдёт точную фразу
- *...гъа* найдёт все слова, состоящие из 6 букв и оканчивающиеся на *-гъа*
- заменяют отсутствующий морфологический анализатор

Статистика

- 1988330 символов (букв), 371536 токенов (слов включая пунктуацию), 51965 разных словоформ
- 287804 (77.5%) «истинных» слов, без пунктуации и цифр
- 24110 предложений, средняя длина предложения 15.4 токенов
- Средняя длина слова (не включая пунктуацию и цифры) 6.58 символов (букв).

Статистика

- самые длинные слова в корпусе:
радиоэшиттирювлеримизнинъ,
укъукъкъорчалайыджыларгъа,
укъукъкъорчалайыджыларнен

Самые частотные слова в корпусе

ве	5598
бир	3113
бу	1900
ичюн	1873
де	1853
да	1770
эди	1390
Бу	1365
сонъ	1172
озъ	1058

Самые частотные биграммы в корпусе

я да	307
эр бир	243
бир къач	222
ич бир	212
эм де	171
дегиль де	131
ве онынъ	113
Миллий Меджлис	103
бар эди	98
бир сыра	96

File Edit View Bookmarks Widgets Tools Help

New tab http://korpus.juls.savba.s... x


http://korpus.juls.savba.sk/QIRIM 100%

Query

бала.* Араштырув корпуста: QIRIM

АБВГГЪДЕЁЖЗИЙККЪЛМННЪОПРСТУФХЦЧДЖШЩЪЫЬЭЮЯ

абвггъдеёжзийккълмннъопрстуфхцчджшщъыьэюя

Опциялар • [Инфо](#) • [Корпус акъкында](#) • 

52418	()	Нефакъанынъ экинджи чешити	бала	догъгъан заманда бир
52453	()	язып , чалышкъан ерине	баланынъ	догъумы акъкында шеадетнаме
52466	()	этмек керек . Эки я да зияде	бала	догъгъан олса , нефакъа
52473	()	догъгъан олса , нефакъа эр бир	балагъа	тёлениле . Эгер бала олю
52477	()	балагъа тёлениле . Эгер	бала	олю догъгъан олса , нефакъа
52491	()	чешити учь яшына къадар	баланы	бакъмакъ ичюн ярдымдыр
52531	()	шу чалышкъан къадынгъа	баласы	учь яшына толгъангъа
52572	()	Эльбетте , аиледе къач	бала	оладжагъы шахсий иш .
52669	(x)	себебини сорасам , онынъ	баласы	яныккъа огърагъан экен
52692	()	башладым . Бираздан сонъ	бала	тюзельди , яныкъ ярасындан

? ” деген суаллериме багъчасарайлы Айдер Керимов бойле джевапланды : “ Мен ватандашларымыз киби эв къурам . Бир кунь менде чалышкъан къуруджылардан бири кельмеди . Кельмегенининъ себебини сорасам , онынъ **баласы** яныккъа огърагъан экен . Чокъ тюшюнмейип , онынъ эвине бардым . Балагъа аджеле ярдым бердим , оны тедавийлемеге башладым . Бираздан сонъ бала тюзельди , яныкъ ярасындан ич бир

190..199/915 (10)

[биринджиси](#) [эвелькиси](#) [невбеттеки](#) [сонъкиси](#)

Query

бала.*

Араштырув корпуста: QIRIM

АБВГГЪДЕЁЖЗИЙКЪЛМННЪОПРСТУФХЦЧДЖШЩЪЫЬЭЮЯ

абвггъдеёжзийккълмннъопрстуфхцчджшщъыьэюя

[Опциялар](#) • [Инфо](#) • [Корпус акъкъында](#) •    

87160	()	иштирак эткенлернинъ	балалары	, бу адамлар къайсы бир
87388	()	чаршанбе куню участка	бала	экими чалыша . Ана - бабалар
87396	()	чалыша . Ана - бабалар озь	балаларны	хастаханеге (поликлиника
88120	()	Пакизе бизге (биз беш	бала	эдик) эр вакъыт : « Балалар
88195	()	этмеге севе эди . Биз ,	балалар	, ве эвимизге кельген
88424	()	Ускут коюнден . Аиледе беш	бала	тербиелендик . Учъ огълан
88535	()	тапалмагъан сонъ , къадынымны ве	балаларны	къайнатамнынъ эвинде
88918	()	Биринджиден , медреседе окъугъан	балалар	16 - 19 яшында мусульман
88967	()	. Тек бир проблем бар :	балалар	ана тилини зайыф билелер
89991	(x)	сенеси эвини сатып , дёрт	балалы	аиле ана - Ватангъа авдет

бите , кери – аилесине къайтмагъа меджбур ола . Гъурбетте кене авдет ичюн куреште фааллик эте . Иши ве къазанчы яхшы олгъанына бакъмай , 1977 сенеси эвини сатып , дёрт **балалы** аиле ана - Ватангъа авдет олып , Къарасувбазар районынынъ дагъ этегиндеки дюльбер манзаралы койде , халкъымызнынъ эфсаневий къараманы Алим Азамат огълу догъыпоськен Копюрликой топрагъында ерлеше . Аиле башы дёрт

280..289/915 (10)

[биринджиси](#) [эвелькиси](#) [невбеттеки](#) [сонъкиси](#)

Query

бала.*

Араштырув корпуста: QIRIM

АБВГГЪДЕЁЖЗИЙККЪЛМННЪОПРСТУФХЦЧДЖШЩЪЫЬЭЮЯ

абвггъдеёжзийккълмннъопрстуфхцчджшщъыьэюя

[Опциялар](#) • [Инфо](#) • [Корпус акъкъында](#) •    

94083 ()	ичильген беш фильджан чай	бала	инкишафына зарар эте
94184 (x)	анълаймыз . Бала не къадар	балабан	олса , о къадар онынъ вуджудыны
94197 ()	теминлемек агъыр . Эксериетте	балалар	базы бир махсулатларны
94222 ()	. Чокъ витамин ашагъан	балалар	акъылланырлармы ? Витаминлерни
94230 ()	фајдаланув орта севиели	баланы	Энштейнге чевирмез . Бунынънен
94242 ()	мадделернинъ етишмеюви	баланынъ	зийреклигине , бирде -
94274 ()	вуджудда демирнинъ етишмеюви	баланынъ	табиатына ве ишке авеслигине
94302 ()	азырлады . Сентябрь 2005Энъ	балабан	той Сакъта олды Къырымларнынъ
95401 ()	Бу девирде ана - бабалар	балаларына	паалы чалгъы алетлери
99317 ()	талебелери рамазан айында	балалар	эв - эвден , къапу - къапудан

алмашув деген шейни ерине кетирип догъру ёлгъа къоялар . Олар осювге , яхшы кейфке ве сагълыкъкъа тесир этелер . Оларнынъ муим ролюни тек олар етишмегенинде анълаймыз . Бала не къадар **балабан** олса , о къадар онынъ вуджудыны муим гъыдаларнен теминлемек агъыр . Эксериетте балалар базы бир махсулатларны ред этип дигер , сеvimли махсулатларны талап этелер . Шунынъ ичюн вуджудгъа яшайышкъа

300..309/915 (10)

биринджиси

эвелькиси

невбеттеки

сонъкиси