

Turklang 2021

The Creation of the Kyrgyz Corpus

Kyrgyz-Turkish “Manas” University

PhD Aida Kasieva

Bishkek, Kyrgyzstan



Availability

The Kyrgyz Corpus is the joint project between the Universität des Saarlandes and the Kyrgyz-Turkish 'Manas' University.

Archived at the CLARIN-D centre in Saarbrücken

CMDI metadata for harvesting

Covered by the Virtual Language Observatory (VLO) CQPweb interface at Universität des Saarlandes





Corpus Composition

novels (3 texts, 563000 tokens)
epic (1 text, 330000 tokens)
novelette (1 text, 20000 tokens)
fairy tale (74 texts, 71000 tokens)
minor epic (5 texts, 221000 tokens)

Kyrgyz Corpus Design

The Universal Copyright Convention became effective in the Soviet Union on 27 May 1973. Before that date, only local copyright laws existed in the Soviet Union. This has the effect that literary works created and published before the year 1972 in the Soviet Union are in the public domain now. This made possible to use the literary works from the website Bizdin.kg, which hosts them.



Corpus Data

The data used in the corpus are released under a license that acts in accordance with Kyrgyz laws on distribution.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details:

<http://creativecommons.org/licenses/by/4.0/>

1. Source: <https://www.ethnologue.com/18/language/kir/>
2. http://corpora.uni-leipzig.de/en?corpusId=kir_community_2017
3. <https://www.clarin.eu/>
4. <https://vlo.clarin.eu>
5. <https://www.clarin.eu/content/content-search>
6. <https://www.apertium.org/>



Text documents description (Metadata: Title, Source, Author)

The corpus title is: The Kyrgyz Corpus (2019-04-18): powered by CQPweb. Every document contained in the corpus is stored in a plain text format in the UTF-8 encoding.

In the Menu section there is a subsection Corpus Info that comprises Corpus Metadata.

The main page of the Kyrgyz corpus

Kyrgyz Corpus (2019-04-18): powered by CQPweb

Metadata for Kyrgyz Corpus (2019-04-18)

Corpus title	Kyrgyz Corpus (2019-04-18)
CQPweb's short handles for this corpus	kyrgyz_20190418 / KYRGYZ_20190418
Total number of corpus texts	84
Total words in all corpus texts	1,243,161
Word types in the corpus	92,263
Standardised type:token ratio (1,000-token basis)	Cannot be displayed (STTR not cached)
Non-standardised type:token ratio	0.0742 types per token

Text metadata and word-level annotation

The database stores the following information for each text in the corpus:	There is no text-level metadata for this corpus.
The primary classification of texts is based on:	A primary classification scheme for texts has not been set.
Words in this corpus are annotated with:	PoS
	stem
The primary word-level annotation scheme is:	PoS



Kyrgyz Corpus Structure

Corpus title – the title of a document;

CQP web's short handles for this corpus;

Total number of corpus texts;

Total words in all corpus texts;

Word types in the corpus;

Type: token ratio

S-attributes in this corpus:

<code><s></code>	Structure ``s''
<code><text></code>	Structure ``text''
<code><text_author></code>	Structure ``text_author''
<code><text_genre></code>	Structure ``text_genre''
<code><text_id></code>	Structure ``text_id''
<code><text_title></code>	Structure ``text_title''
<code><text_title_english></code>	Structure ``text_title_english''
<code><text_year></code>	Structure ``text_year''

The text is tokenized by an in-house tool and lemmatized and POS-tagged using the Apertium toolkit (Washington et al. 2012). For convenient use, the corpus is post-processed to a vertical format as used by the Corpus Query Processor and CQPweb (Hardie, 2012).

Tokenization and annotation of the corpus data

```
724689 ^берди/бер<vaux><ifi><p3><sg>$
724690 ^,/,<cm>$
724691 ^тагыраак/так<adj><comp>$
724692 ^айтканда/айт<v><iv><ger_past><loc>$
724693 ^,/,<cm>$
724694 ^аны/ал<prn><pers><p3><sg><acc>$
724695 ^адаштырышты/адаш<v><iv><caus><ger_pres><acc>$
724696 ^,/,<cm>$
724697 ^ишкердик/ишкердик<n><nom>$
724698 ^туткунуна/туткун<n><px3sp><dat>$
724699 ^алышты/ал<v><tv><coop><ifi><p3><pl>$
724700 ^,/,<cm>$
724701 ^бирок/бирок<snjcoo>$
724702 ^азыр/азыр<adv>$
724703 ^кеп/кеп<n><nom>$
724704 ^ал/ал<prn><pers><p3><sg><nom>$
724705 ^жөнүндө/жөн<n><px3sp><loc>$
724706 ^эмес/э<cop><neg><aor><p3><pl>$^)/<rpar>$
724707 ^./.<sent>$
724708 <^///<sent>$^s/*s$>
724709 <^s/*s$>
724710 ^Бул/бул<det><dem>$
724711 ^элес/элес<n><nom>$
724712 ^анын/ал<prn><dem><gen>$
724713 ^жетимсиреген/*жетимсиреген$
724714 ^жүрөгүндө/жүрөк<n><px3sp><loc>$
724715 ^чыныгы/чыныгы<adj>$
```

Selection of Tagsets


- Apertium6, an open-source machine translation platform has been used for Kyrgyz corpus annotation (POS tagger).
- Tokenisation, creation of a vertical format (vrt)
- Part-of-speech tagging using Apertium
- Apertium is based on the Helsinki Final State Transducer
- Artefact removal

Your query "[word='Манас.*']" returned 3,353 matches in 8 different texts (in 1,243,161 words [84 texts]; frequency: 2,697.16 instances per million words) [0.024 seconds - retrieved from cache]

|< << >> >| Show Page: 1 Line View Show in random order Choose action... Go!

No.	Text	Score
1	Manas01	
2	Manas01	37 - 9 Кыргыз элинин улуу мурасы , улутту
3	Manas01	көөдөнүнө уюткудай кармап , байыркы кыргыздын обондуу [UNREADABLE] поэзиясын С
4	Manas01	менен уккусу келген замандаштарыбыз көп эле . Де
5	Manas01	негизги парзына , жашоосунун маңызына айланганы бекер
6	Manas01	улуттун дүйнө таануусун , тарыхый ордун , ишени
7	Manas01	эпикалык инсан . Ушундайча айтууга мүмкүн болсо , биз мамлекетт
8	Manas01	өзгөрүп турган жандуу жайдары мүнөз , ачык ,
9	Manas01	Мунун баарын текшилеп айтып жатканым , башында берген у
10	Manas01	менен жолугуп калганда кошуна иретинде Саякбай Каралаевди чакырып с
11	Manas01	туудуруп , айрыкча суктандыраар эле . Мухтар Ауэзов менен Дм
12	Manas01	И-й б...


Kyrgyz Corpus as a Poster



CLARIN-D
FEDERAL MINISTRY OF EDUCATION AND RESEARCH


A NEW KYRGYZ CORPUS: SAMPLING, COMPILATION, ANNOTATION

AIDA KASIEVA, JÖRG KNAPPEN, STEFAN FISCHER & ELKE TEICH
aida.kasieva@manas.edu.kg, {j.knappen, e.teich}@mx.uni-saarland.de,
stefan.fischer@uni-saarland.de



UNIVERSITÄT DES SAARLANDES

THE KYRGYZ LANGUAGE



- Official language of Kyrgyzstan
- Turkic language, Kypchak branch

CQPWEB DEMO

Query "[word="Манас.*"]" returned 3,353 matches in 8 different texts (in 1,243,161 words [84 texts]; frequency: 2,697.16 instances per million words)

Showing frequency breakdown of words in this query, at the query node: there are 34 different types and 3,353 tokens at this concordance position. [1,314 words]

|< << >> >| Breakdown position: Node Frequency breakdown of words only Go!

No.	Query result	No. of occurrences	Percent
1	Манас	2199	65.58%
2	Манастын	328	9.78%
3	Манасты	313	9.33%
4	Манаска	267	7.96%
5	Манастан	60	1.79%
6	Манастар	50	1.49%
7	Манастар	33	0.98%
8	Манастар	25	0.75%
9	Манасты	15	0.45%
10	Манастар	10	0.3%
11	Манастар	9	0.27%
12	Манасты	8	0.24%
13	Манастары	6	0.18%

https://corpora.clarin-d.uni-saarland.de/cqpweb/kyrgyz_20190418/

Thank You for your attention!

