A new Kyrgyz corpus: sampling, compilation, annotation

Mittwoch,
04.03.2020
15:45–16:30
ESA1 W Foyer

Aida Kasieva[1] , Jörg Knappen[2], Stefan Fischer[2], Elke Teich[2]

[1] *Kyrgyz-Turkish Manas University, Bishkek, Kyrgyzstan,* [2] *Universität des Saarlandes, Saarbrücken*

aida_kasieva@yahoo.com, j.knappen@mx.uni-saarland.de, e.teich@mx.uni-saarland.de, stefan.fischer@uni-saarland.de

CLP

We present a Kyrgyz corpus comprising 1,205,888 words of 84 literary texts of five genres: novel, novelette, epic, minor epic, and fairy tale. The corpus is annotated with part-of-speech tags and rich per-text meta-data and made available under a free licence from CLARIN-D.

The Kyrgyz language is spoken by approximately 4.4 million speakers worldwide and it is one of the two official languages of Kyrgyzstan. It belongs to the Turkic language family and has a rich agglutinative morphology. Kyrgyz is currently a low-resourced language, available corpora include web corpora without further annotation* and a plain text Wikipedia dump**. This corpus is to our knowledge the first fair-sized Kyrgyz corpus containing human-curated texts with lemma and part-of-speech annotations.

Corpus sources. As a copyright law became effective in the then Soviet Union only on May 27, 1973, all texts created before that date are public domain. This allowed us to include "Jamila" by the famous Kyrgyz writer Chinghiz Aitmatov in the corpus. Other sources included in the corpus were taken from http://bizdin.kg run by the Bizdin Muras foundation which promotes the development of the Kyrgyz language under a Creative Commons licence.

Meta-data. Each text in the corpus is annotated for the year of publication, author, title, an English translation of the title, and text source. The corpus as a whole is described by extensive metadata in Dublin Core and CMDI (Broeder et al. 2011) format.

Corpus processing. 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).

* Kyrgyz corpus in the Leipzig corpora collection http://corpora.uni-leipzig.de/en?corpusId=kir_community_2017 and kyWaC – Kyrgyz corpus https://www.sketchengine.eu/kywac-kyrgyz-corpus/

** http://hdl.handle.net/11234/1-2735

References: Broeder, D, O. Schonefeld, T. Trippel, D. Van Uytvanck & A. Witt (2011). A pragmatic approach to XML interoperability – the Component Metadata Infrastructure (CMDI). In Balisage: The Markup Conference, vol. 7. Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. International Journal of Corpus Linguistics, 17, 380–409. Washington, J. N., M. Ipasov & F. M. Tyers (2012). A finite-state morphological transducer for Kyrgyz. In Nicoletta Calzolari (Conference Chair) et al. (ed.), Proceedings of the Eighth Conference on Language Resources and Evaluation, LREC2012, Istanbul, Turkey, 934–940.