Corpus Linguistics:

Studies in Crimean Tatar Language

Lenara Kubedinova
Crimean Industrial Pedagogical University,
Crimea
kubedinova@gmail.com
l.kubedinova@mail.ru

Radovan Garabík
Ľ. Štúr Institute of Linguistics,
Slovak Academy of Sciences,
Bratislava, Slovakia
garabik@kassiopeia.juls.savba.sk

Abstract

This paper presents the first steps of Crimean Tatar language in the field of corpus and computer linguistics. First of all it is the current situation of the first corpus of the Crimean Tatar language – "the Linguistic corpus of Crimean Tatar language". Secondly, the creating of Corpus of Crimean Tatar Wikipedia. Thirdly, the prospects of foundation of the joint research laboratory of Turkic computer linguistics which aims to support the functioning of Crimean Tatar language in information communicative technologies.

Key words

Crimean Tatar language, corpus linguistics, Wikipedia.

Introduction

Corpus linguistics is a new, quick developing field in linguistics. This field of computational linguistics deals with the development of main principles of building and using linguistic corpuses with the help of computer technologies. Nowadays every language needs to have a corpus of texts, especially insufficiently known languages. The Crimean Tatar language is listed as "severely endangered" in the UNESCO *Atlas of the World's Disappearing Languages.* The Crimean Tatar language does not dispose of its own electronic library or text database owing to that it has an acute need in corpus linguistics studies.

1. Linguistic Corpus of the Crimean Tatar language

The first corpus of electronic texts of the Crimean Tatar language was created in 2006 (http://korpus.juls.savba.sk/QIRIM/). The authors of this corpus are Lenara Kubedinova, an associate professor of English Philology, Taurida National University, Crimea and Radovan Garabik, a researcher at the Institute of Linguistics named after L. Stuhr (Bratislava, Slovakia). Linguistic corpus of the Crimean Tatar language is aimed at creating a database of contemporary written language. At the present moment it includes newspaper articles from the Crimean Tatar newspapers "Kırım" and "Yanı dunya", several books of Crimean Tatar writers, as well as some poetry of the Crimean Tatar poets. At the beginning the corpus included the texts only in Cyrillic script. The work on supplementing the corpus with the texts in Crimean Tatar language was suspended for a couple of years view to different reasons (basically financial ones), in spite of that, linguists have always had free access to it.

Currently, the Linguistic corpus of the Crimean Tatar language is growing, due to the supplement with texts in Cyrillic, as well as creating the corpus of texts in Latin. The texts in Latin are taken from the Crimean

Tatar newspaper "Yanı dunya", as well as women's literature and art, popular science magazine "Nenkecan."

The corpus is provided free of charge to all users without the need of registration. NoSketch Engine (Rychlý 2007) is used as a user interface system, which allows a simple search of a word or phrases (some words in a regular order), or arbitrary regular expressions (which allows in a certain degree to replace not existing morphological analysis and lemmatization) on the level of words and phrases, and provides statistical analysis on different criteria. You can be connected to the system by an ordinary Web browser; the old system (Manatee1 / Bonito) also works, but it isn't developing, and we offer all users to use the new system.

The current state of the Crimean Tatar corpus in Cyrillic is: 6015978 symbols (letters), 521754 tokens (including punctuation), that makes up 65429 of different wordforms. 405092 «real» words, without punctuation and numbers (77.6%).

2. Corpus of Crimean Tatar Wikipedia

Corpus of Crimean Tatar Wikipedia was created in September 2014. Crimean Tatar wikipedia (in Latin script) officially started on 12 January 2008, though the version in the incubator dates from September 2006. At the time of writing, it contains about 4000 articles, which makes it one of the very small wikipedias available, ranking 164th by the number of articles. Nevertheless, it represents contemporary living language and as such as a valuable source of texts. We describe the technical conversion process used to obtain the texts and a simple metadata annotation. This subcorpus is released under the same license as the wikipedia itself, i.e. Creative Commons Attribution-ShareAlike 3.0 Unported.

2.1 Text Extraction

All the wikipedia data are publicly available in the form of either raw database dumps, or XML wrapped raw article text data. For our purposes, the XML data was preferred, because it offers easy access to article metadata(timestamp of last edition, title), and it contains the same source text inside.

Wikipedia articles are separated into several namespaces, according to the role of the article in the wiki. The most important (for our purposes) are the main namespace, used for the articles proper (as well as disambiguation pages and redirects), and the *Talk:* namespace[1], used for discussion about articles. Other potentially useful namespaces are *User:* for user pages and *Talk user:* for user discussion – in the Crimean Tatar Wikipedia the number of these is negligible compared to the amount of articles.

2.2 Text Filtering

Since the Wikipedia's content by its very nature is of rather different nature than in "usual" written language, even compared to a traditional encyclopedia, a question of text selection becomes quite important. This is especially striking when considering non-textual information included in the articles, such as templates, infoboxes, pictures and other elements (with their mark-up structures). Wikipedia texts also contain unusual amount of geographical and personal names in their original languages and original scripts (thanks to the ease of collecting such an information thanks to Internet and sister language mutations), something that traditional encyclopaedias have not achieved yet. The discussion pages often contains texts in other languages, apart from English (the working language of the whole Wikipedia), but due to its specific sociopolitical status, also in Russian or Ukrainian.

First, in order to reduce the amount of non-Crimean Tatar texts in the resulting corpus, we investigated some simple statistic that could be useful in determining if the text is in the language. We introduce two parameters, the first one is the ratio of characters of Crimean Tatar language, i.e. the basic Latin alphabet and the characters ñ, ı, ğ, â, ç, ş, ö, ü – we will further refer to the

[1] *Muzakere* in Crimean Tatar Wikipedia

set of these characters as “crh characters” – to the amount of characters of the text:

$$q_1 = \frac{\sum_{x \in \{abc \ldots z ı ñ ğ â\}} x}{\sum x}$$

and the ratio of just the specific characters – further referred to as “unique crh characters” to the text:

$$q_2 = \frac{\sum_{x \in \{ı ñ ğ â\}} x}{\sum x}$$

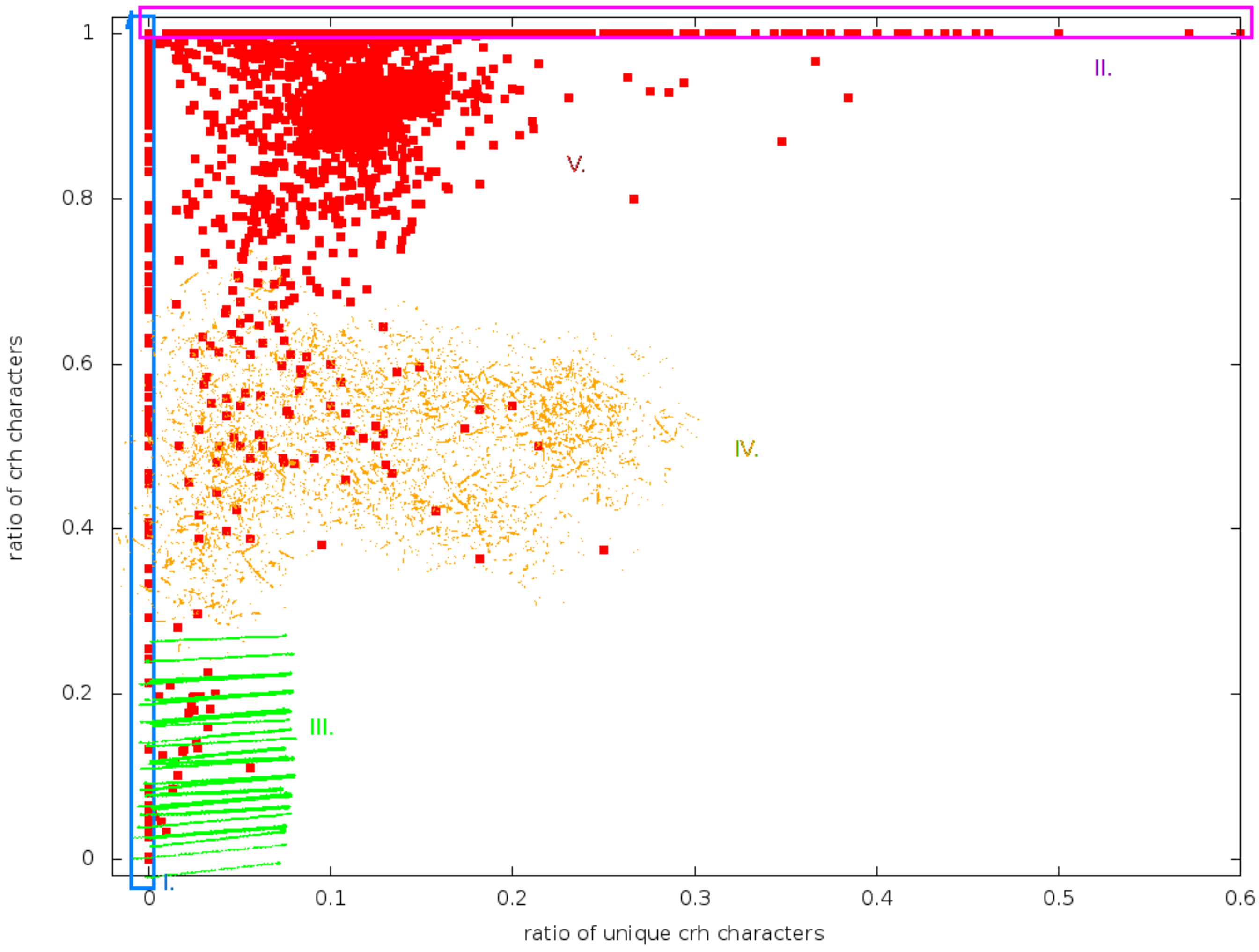


Figure 1: q_1 (y axis) vs. q_2 (x axis) relation. Each dot corresponds to one paragraph of text.

On Figure 1., we can identify several regions of interest:

I. the vertical line q_1=0 – texts, where there are no unique crh characters present

II. the horizontal line q_2=1 – texts, where all the characters are ASCII only

III. the region of low q_1 and low q_2 – basically, the text has some unique crh and ASCII characters, but their ratio to the rest is low; these are mostly texts in Cyrillic with some words in (Latin script) Crimean Tatar

IV. the region of medium ratio of crh characters, nonzero ratio of – these are mostly bilingual texts (such as geographical names with the name in the original language), equations with an explanation etc.

V. the region of high ratio of crh character and nonzero ratio of unique crh characters – these are the “normal” texts, most desirable to remain in the corpus

There is a notable gap between the regions I and the rest of the texts, which allows us to discard this region completely. In this way, we are discarding some false negatives as well, but these are comprised predominantly out of very short sentences (typically one word or two), where by chance no unique crh character appeared. We also discard the region III. Inclusion of region IV is debatable – from a purely corpus linguistic point of view, this contains too much extralingual texts that distorts the statistics eventually obtained from the corpus, but on the other hand, given the small amount of data in the corpus, we decided to provisionally keep the region IV in, at least until we get more experience from using the corpus. Therefore, our final condition for texts retained in the corpus is:

$$q_1>0 \wedge q_2>0.3$$

Overall, the size of the input XML file is 17486151 characters[2], the size of namespaces 0 and 1 before filtering: 15111777 characters, after the first filter 2333809 characters and after the second filter 2127032 characters.

2.3 Metadata Formatting

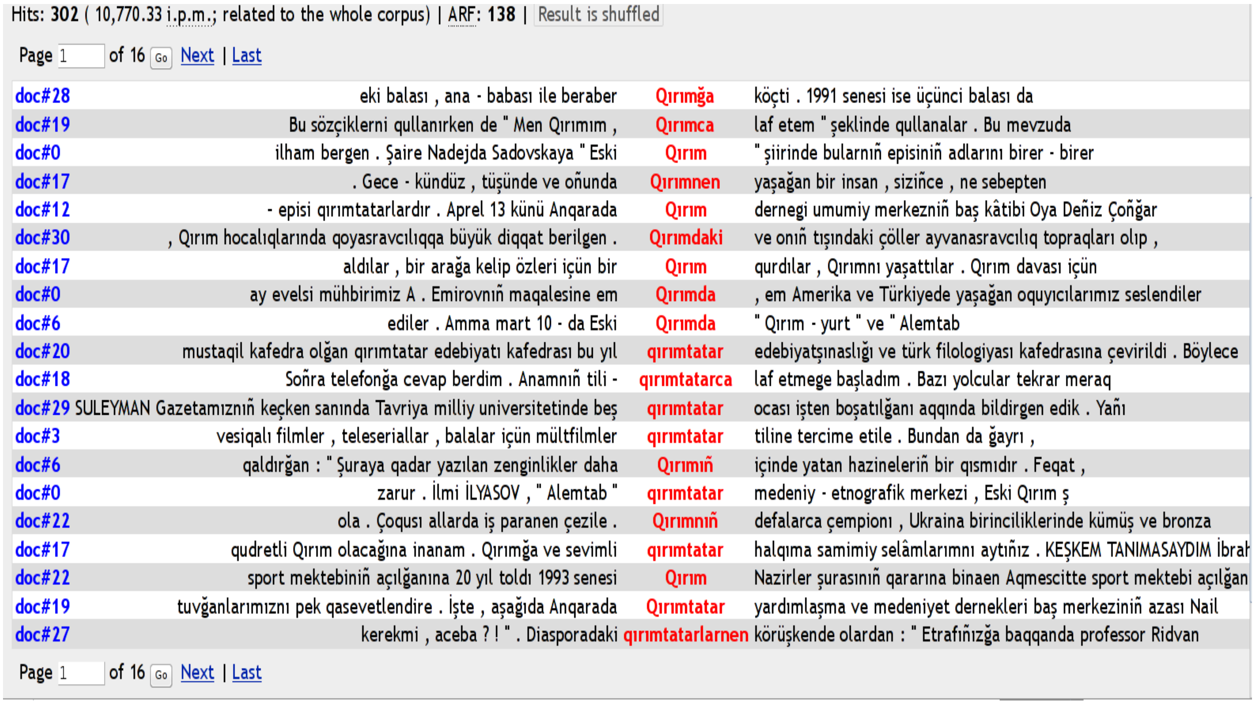
The corpus is logically divided in documents, each document corresponding to one wikipedia article. The metadata describing each document contains timestamp, title and an identification string. In order to facilitate reading corpus query results, we create and identification number by abbreviating document title and adding a unique number (index number of the article). This allows the users to see immediately whether (or how frequently) the term occurs in one document.

```
<doc id="Qırımt~504" timestamp="2014-03-23T02:06:54Z"
title="Qırımtatartamğası">
```

Table 1: Example of metadata annotation in the corpus

In Table 1 we see an example of metadata annotation – the article titled Qırımtatar tamğası, last edited 23rd March 2014, and a short abbreviated identification string Qırımt~504 (the document being 504th in the Wikipedia).

[2] not bytes – the file is in UTF-8 encoding



Hits: **302** (10,770.33 i.p.m.; related to the whole corpus) | ARF: **138** | Result is shuffled

Page 1 of 16 Go Next | Last

doc#28	eki balası , ana - babası ile beraber	Qırımğa	köçti . 1991 senesi ise üçünci balası da
doc#19	Bu sözçiklerni qullanırken de " Men Qırımım ,	Qırımca	laf etem " şeklinde qullanalar . Bu mevzuda
doc#0	ilham bergen . Şaire Nadejda Sadovskaya " Eski	Qırım	" şiirinde bularnıñ episiniñ adlarını birer - birer
doc#17	. Gece - kündüz , tüşünde ve oñunda	Qırımnen	yaşağan bir insan , siziñce , ne sebepten
doc#12	- episi qırımtatarlardır . Aprel 13 künü Anqarada	Qırım	dernegi umumiy merkezniñ baş kâtibi Oya Deñiz Çoñğar
doc#30	, Qırım hocalıqlarında qoyasravcılıqqa büyük diqqat berilgen .	Qırımdaki	ve onıñ tışındaki çöller ayvanasravcılıq topraqları olıp ,
doc#17	aldılar , bir arağa kelip özleri içün bir	Qırım	qurdılar , Qırımnı yaşattılar . Qırım davası içün
doc#0	ay evelsi mühbirimiz A . Emirovnıñ maqalesine em	Qırımda	, em Amerika ve Türkiyede yaşağan oquyıcılarımız seslendiler
doc#6	ediler . Amma mart 10 - da Eski	Qırımda	" Qırım - yurt " ve " Alemtab
doc#20	mustaqil kafedra olğan qırımtatar edebiyatı kafedrası bu yıl	qırımtatar	edebiyatşınaslığı ve türk filologiyası kafedrasına çevirildi . Böylece
doc#18	Soñra telefonğa cevap berdim . Anamnıñ tili -	qırımtatarca	laf etmege başladım . Bazı yolcular tekrar meraq
doc#29	SULEYMAN Gazetamıznıñ keçken sanında Tavriya milliy universitetinde beş	qırımtatar	ocası işten boşatılğanı aqqında bildirgen edik . Yañı
doc#3	vesiqalı filmler , teleseriallar , balalar içün mültfilmler	qırımtatar	tiline tercime etile . Bundan da ğayrı ,
doc#6	qaldırğan : " Şuraya qadar yazılan zenginlikler daha	Qırımıñ	içinde yatan hazineleriñ bir qısmıdır . Feqat ,
doc#0	zarur . İlmi İLYASOV , " Alemtab "	qırımtatar	medeniy - etnografik merkezi , Eski Qırım ş
doc#22	ola . Çoqusı allarda iş paranen çezile .	Qırımnıñ	defalarca çempionı , Ukraina birinciliklerinde kümüş ve bronza
doc#17	qudretli Qırım olacağına inanam . Qırımğa ve sevimli	qırımtatar	halqıma samimiy selâmlarımnı aytıñız . KEŞKEM TANIMASAYDIM İbrah
doc#22	sport mektebiniñ açılğanına 20 yıl toldı 1993 senesi	Qırım	Nazirler şurasınıñ qararına binaen Aqmescitte sport mektebi açılğan
doc#19	tuvğanlarımıznı pek qasevetlendire . İşte , aşağıda Anqarada	Qırımtatar	yardımlaşma ve medeniyet dernekleri baş merkeziniñ azası Nail
doc#27	kerekmi , aceba ? ! " . Diasporadaki	qırımtatarlarnen	körüşkende olardan : " Etrafıñızğa baqqanda professor Ridvan

Page 1 of 16 Go Next | Last

*Figure 2: Example of corpus interface, query qırım.**

3. Prospects

Joint Research Laboratory of Turkic computer linguistics

The work on the Linguistic corpus of the Crimean Tatar language was proceeding quite slowly because it had been done voluntarily only by two persons (one programmer and one philologist) without any financial support. In summer 2014 the Institute of Applied Semiotics of the Academy of Sciences of Tatarstan offered to the Faculty of Crimean Tatar and Turkish Philology of the Crimean Industrial Pedagogical University establish a Joint Research Laboratory of Turkic computer linguistics. The idea was supported by the President of Tatarstan Rustam Minnekhanov. This fact gives a great opportunity not only to develop the first corpus of the Crimean Tatar language but to provide the functioning of the Crimean Tatar language in information technologies.

Presumably the laboratory starts its work from January 2015.

A wide range of tasks are set:

- Carrying out of the fundamental scientific and applied researches in the field of computer linguistics in order to solve problems on computer supply of functioning the Crimean Tatar language;
- Development of formal models used to describe the Crimean Tatar language;
- Creation of software development for the automated processing of the Crimean Tatar language:

particularly:

- the development of the Linguistic corpus of the Crimean Tatar language
- morphological analyzer of the Crimean Tatar language
- morphological corrector of the Crimean Tatar language, synthesizer and recognizer of the Crimean Tatar language
- Tatar – Crimean Tatar machine translator;

- Electronic atlas of the Crimean Tatar dialects, etc.

- Implementation of the Crimean Tatar language in software products of leading firms: ABBYY (Lingvo and FineReader), Microsoft and mobile devices.
- Development of training courses using scientific and applied the results obtained in the framework of the research activity of the Laboratory.
- Create optimal conditions for joint basic and applied research, conducted by the University and the Academy in the various scientific programs, projects and grants, including international, as well as the expansion of scientific and industrial contacts with other academic institutions.
- Organization of training courses for young scientists at the University in the leading research units of the Academy and international research centers.
- Involvement of leading scientists of the Academy for giving special courses for students of the Faculty of Crimean Tatar and Turkish Philology of the University, scientific management of graduates and consulting of doctoral candidates
- Organization and carrying out of scientific seminars and conferences.
- Preparation of joint scientific articles, monographs and textbooks.

Such laboratory offers an opportunity to bring the development of the Crimean Tatar language on a quite new technological level that gives a great hope to resuscitate this rich and beautiful language.

References

Rychlý, Pavel. Manatee/Bonito – A Modular Corpus Manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, 2007. p. 65–70. ISBN 978-80-210-4471-5.

Кубединова Л.Ш., Р.Гарабик Лингвистический корпус крымскотатарского языка: перспективы развития // Труды Казанской школы по компьютерной и когнитивной лингвистике, TEL-2014 (Казань, 6-9 февраля 2014г.), Выпуск 16. – Казань, 2014. – С. 124-127.