Oral Tradition, 28/2 (2013):193-204

On the Principles of a Digital Text Corpus: New Opportunities in Working on Heroic Epics of the Shors

Dmitri A. Funk

Introduction

The Corpus of Folklore Texts in the Languages of Indigenous Peoples of Siberia (http://corpora.iea.ras.ru) is a collection of folklore texts in lesser used, mostly endangered, Siberian languages and was initiated in 2011 through support from the Department of Northern and Siberian Studies at the Institute of Ethnology and Anthropology of the Russian Academy of Sciences (RAS).[1] The main idea was to create a corpus capable of storing both an original version and an orthographically standardized version of any text. Another task was to build a system able to perform some related analytical procedures. At the same time we were aiming at making a significant part of the unknown materials available for researchers and others able to read in the native Siberian languages.

This essay will use the Shor corpus as its main example, though the Teleut, Evenki, or Nenets corpora might easily have been chosen instead. The reasoning behind this choice is the fact that the whole project arose out of my long-term study of the Shor epics. Additionally, the Shor (like the Teleut) materials belong for the most part to me, and I am the primary individual who has been working with them within the frame of this project. It is my hope that this single example will help readers better understand what our Corpus is able to do and how it has been organized.

[1] I am thankful to the RAS Presidium Corpus Linguistics research initiative (which in its initial phase was headed up by myself until my former Ph.D. student and colleague Dr. Kyrill Shakhovtsov took the lead in 2012) and to the Foundation for Fundamental Linguistic Research (Project A-16-2013) for their support of the Corpus and related projects. My thanks go also to the Institute of Education of Indigenous Peoples of the North, Siberia, and the Far East of the Russian Academy of Education and to the National Research Tomsk State University (in relation to the "Man in a Changing World: Problems of Identity and Social Adaptation in History and at Present" project; RF Government grant No. 14.B25.31.0009) for their assistance that allowed me to focus my attention on working with South Siberian ethnographic data.

The Shors and Their Epics

The Shors are one of Siberia's smaller populations: 12,888 according to the 2010 Russian census.[2] This ethnic group lives mainly in the south of Western Siberia. Until the early twentieth century they provided a perfect example for illustrating the cultural way of life of taiga hunters, gatherers, and fishermen; they now primarily occupy towns in the southern part of the Kemerovo region, where we currently find about 73% of the entire Shor population. In the last 24 years not only have the Shors dwindled in number, but the percentage of those who command the Shor language has fallen as well. Only a small percentage of the Shor—around 5 to 10% (officially 22%)—are still able to speak Shor as their mother tongue, making it an extremely endangered Turkic language.[3]

This ethnic group is especially well known thanks to its rich heroic epic tradition, examples of which have been recorded over the last 150 years by Wilhelm Radloff, Alexander V. Adrianov, Nadezhda P. Dyrenkova, Georgij F. Babushkin, Alexander I. Smerdov, Olga I. Blagoveshchenskaya, and Andrei I. Čudoyakov, as well as by other scholars, enthusiasts, and even some story-tellers.[4] Like epic tales of many other Turkic-speaking groups of Siberia—and as is also the case for Mongolian heroic epic tales—Shor epics most often involve either one or both of two themes: one where the hero embarks upon a quest for a wife and another where the hero struggles with foreign invaders. But these two main themes can be and still are realized in hundreds of variations or motif-series, using the whole richness of an epic tradition.[5]

Fig. 1. Main territories now occupied by the Shors (originally published in Funk and Tomilov 2006:247).

I point especially at heroic epics because they are of special value in studying the language;[6] indeed it is through the language of the heroic epics—and this fact has been long ascertained by linguists—that the Shor language achieved its "higher," one could say "literary," form. It is in the heroic epics that we discover the richness both of the lexical makeup of the language and of its grammatical structure, with every form from the simplest syntax to the most complex being featured within these works.

---

[2] Basic data about the Shors is taken from Funk and Tomilov 2006.

[3] Russian linguists relate the Shor language to the Khakass subgroup of the Uygur-Oguz group within the Eastern-Hun branch of the Turkic languages (Baskakov 1981:20), although one can also consider both the Mras and Kondoma dialects of the Shor to be equally close to Khakass and Teleut respectively, that is, to languages that belong to different groups of the Turkic languages.

[4] For further details see Funk 2005:204-51 and 2010-13.

[5] See Čudoyakov 1995; Arbačakova 2001; Funk 2005.

[6] See, for instance, Funk 1999, 2003, 2006; Esipova and Arbačakova 2006.

Principles of the Corpus

There are five main principles on which the Corpus is based:

1. To increase the number of epic texts available for researchers. There are at least 265 texts of the Shor epics stored in different archives and/or private collections. From these rich materials there were but 26 epic texts published in the original (Shor) language between 1861 and 2010, and only 17 of them were full-length. Most earlier publications by Radloff (1866) and Dyrenkova (1940) are very difficult to access, especially for the Shors in their places of residence and in some cases even for researchers.

On the other hand, the Shor part of the Corpus is the largest one.[7] It now comprises 183,580 orthographic words in 41 texts and is supplemented by an oral sub-corpus of approximately 30,000 words, currently accessible in audio form only.[8]

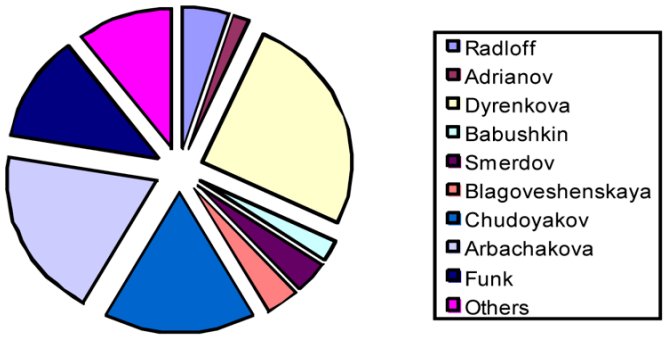


Fig. 2. Proportions of the numbers of the Shor epic texts recorded by scholars between 1861-2006.

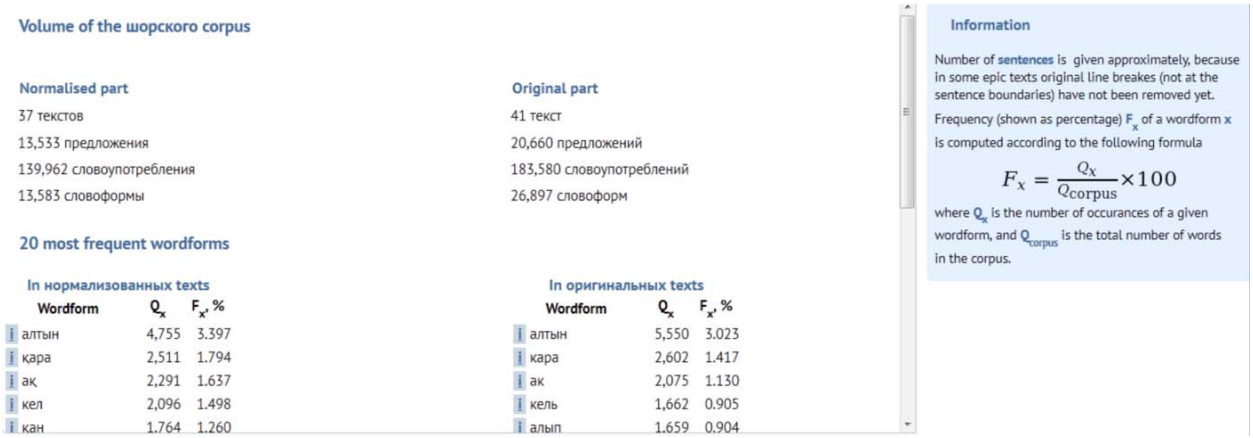


Fig. 3. Volume of the Shor Corpus (with a list of the 20 most frequent word-forms) from http://corpora.iea.ras.ru/corpora/statistics.php.

Out of the 38 texts that represent the Shor heroic epics there are 27 that are unique to this corpus: they were originally part of my own personal collection and (excepting six texts that appeared recently as part of Funk 2010-13) had not been previously published in any form. There is no other freely accessible corpus in Shor or in any other Siberian Turkic language of comparable volume.

[7] All figures here and below reflect the state of the Corpus on October 5, 2013.

[8] The English version of the website is far from being excellent, so here and below there are some obvious gaps and inexact translations in the images.

2. To reflect the very complicated dialectal system of Siberian languages. Our texts represent both dialects of the Shor language. Strictly speaking there are three sub-dialects, including the literary dialect, and in the very near future there will be at least five. Since the Shor still maintain many sub-dialects (not all of which have been well described scientifically) and the literary form is understudied with its many significantly varying "norms," the Corpus will not only help safeguard these sub-dialects and study them, but also contribute to the literary form of the language.

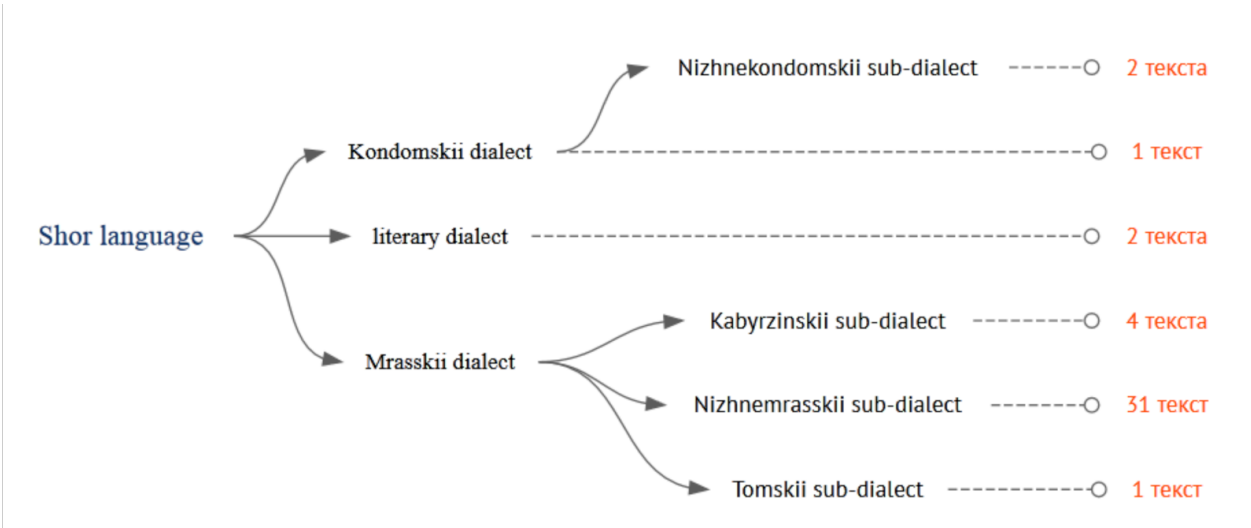


Fig. 4. The structure of the Shor Corpus from http://corpora.iea.ras.ru/corpora/structure.php.

3. To include texts that would give researchers the possibility to analyze the language of folklore on the micro-level. The goal here, then, is to be able to perform analyses on the individual epic idiolects of each singer and—if possible—to do so through a large number of texts. The largest part of the Shor epic texts included within the Corpus derives from the repertoire of the prominent Shor storyteller Vladimir Tannagashev, who left us a large heritage in written form. Tannagashev was born on December 10, 1932, into the family of a hunter. For various reasons, he could not complete general education school studies and began to work early—as he would throughout nearly his entire life—at a coal mine. In the 1940s he started his attempts at narrating epic stories; when he was between 35 and 40, he began to perform them in the guttural singing style. Considering the breadth of his repertoire, Tannagashev may well be considered a singular figure. At 70 he told me that during his younger years he had remembered 122 epics. "Now, having gotten old," he said, he "forgot it all" and could perform only some 70.

We worked constantly with this storyteller from 2001 until the end of 2006; apart from making many records of live performances, I also persuaded Tannagashev to write down his repertoire. As a result, there are 32 epic texts, ranging from 20 to 150 pages each, that were recorded by Vladimir Tannagashev at my request. On the whole, in my estimate, there have been about 40 total recordings, of which 36 are available to me. 25 of these texts have been included in their entirety within the Corpus: 20 texts are Tannagashev's own written versions, and five represent his live performances. If we continue working on the Tannagashev materials, there is a good chance that we will be able to increase the number of his texts to 70-75 since not only did I work with him, but at least one other researcher (L. N. Arbačakova) did as well.

Fig. 5. Vladimir Tannagashev in his apartment (in the kitchen) in the town of Myski, Kemerovo region, 2003. Photo by D. Funk.

Normalised title	Dialectology	Performer	Year
✓ Ай қараттыг Қара-Қан	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Ақ-Пилек	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Алтын-Салғын	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2004
✓ Алып-Кускун	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2000
✓ Ачазы кулатпа туңмазы кулат	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2005
✓ Карағы чоқ Сас-Қараба чодазы чоқ Чол-Қара	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Қара-Қан	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	1999
✓ Қара-Қан	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2002
✓ Қара-Қан	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Көк-Торчуқ	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	1999
✓ Күннӱ көрчең Күн-Көөк	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	1999
✓ Кускун қараттыг Алып-Кускун	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2004
✓ Кырық кулаш сынныг қара сараттыг Алып-Қарачын	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Кырық кулаш сынныг қара сараттыг Қан-Мерген	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Кырық кулаш сынныг қара сараттыг Қара-Молат	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Кырық эмчектиг Қыдай-Арыг	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2003
✓ Кырық эмчектиг Қыдай-Арыг	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2003
✓ Он алыптың ымайынаң чайалған Қан-Кичей	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Сыбазын-Оолақ	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Сывет-Оолақ	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Талашқа чӧрген Алтын-Торғу	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Улуг-кичиг ақ сарат	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Чабыс-Чапан	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2000
✓ Чарық тӱктӱг ақ қалтар аттыг Алтын-Коста	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2006
✓ Чылан-Тоочый	Mrasskii » Nizhnemrasskii	В.Е. Таннагашев	2003

Fig. 6. List of epic texts from the Tannagashev's repertoire in the Shor Corpus from http://corpora.iea.ras.ru/corpora/texts.php?performed_by=17.

Another unique figure on the cultural scene of Mountain Shoriya—and even more broadly in Southern Siberia—during the Soviet era was the Shor storyteller and poet Stepan S. Torbokov. Torbokov was born into the family of a hunter on December 24, 1900, in the Tagdagal *ulus* (Shor settlement) now known as the town Osinniki. He graduated from a parish school at age 12 and was determined to continue his education. However, he did not gain acceptance to the Biisk Catechist College. Instead, in 1927-28 local Soviet authorities sent him to enroll in a month-long course for the preparation of illiteracy eradication specialists in the town of Myski. Having completed it, Torbokov started working at an adult education school, and in 1930, after a short period of training at the Krasnoyarsk Pedagogy School, he became a schoolteacher. Nevertheless, Torbokov continued learning, and he graduated from the Tomsk Institute of Pedagogy in 1943. He retired in 1956 and devoted himself wholly to the mission of recording Shor epics until the end of his life in 1980.

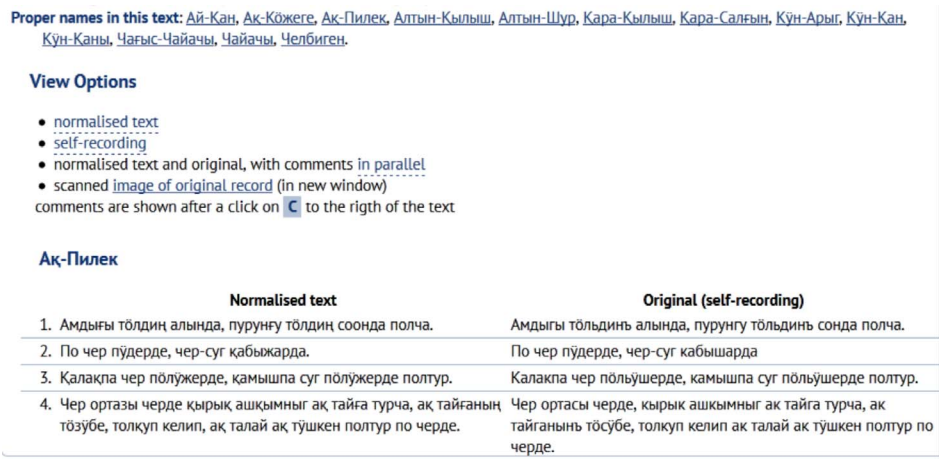
Fig. 7. Taken by an anonymous photographer on June 15, 1969. The picture is accompanied by Torbokov's note in Russian: "(I am) reciting (the epos) *Ak-Salgyn* ["White Wind"] by accompanying on the kay-komus." Stored in the Folklore Archive of the State Literature Museum (No. 419.12. Sheet 1 and originally published in Funk 2010-13).

It was during his youth that Torbokov began to perform epic stories about bogatyrs for his co-villagers. By Torbokov's own estimate, he knew and performed over 40 epics. Already in the 1930s, Torbokov had attempted to write down stories from his repertoire, though the earliest preserved text available today is dated 1941. He further wrote down a couple of stories at the very end of the 1940s, but his main work recording the epics occurred during his retirement in the 1950s and later. The heritage of this outstanding storyteller ended up in various archives, and the work on recovering it is still continuing. In total, the four state archives—in Moscow, Novokuznetsk, Gorno-Altaysk, and Abakan—have preserved for us 40 of these heroic epics. Although every Torbokov text adheres to the bilingual principle—that is, each page of the Shor text is followed by a page of word-for-word Russian translation—unfortunately both the originals and the translations display a rather salient lack of care and also contain numerous stylistic and grammatical errors. In many ways, it is this issue of quality that accounts for the fact that none of these texts has yet been published in an unabridged form. But recently two of his texts have been prepared and one has even been included in its entirety within the Corpus.

It takes a great deal of time and much effort, but in the very near future (depending on time and financial support) I plan to include within the Corpus at least 30 additional epic texts, making them available for the first time. The largest portion will obviously be from Tannagashev's repertoire, but eight of these texts will represent the epic idiolect of another famous story-teller, Maria Tokmagasheva (1908-1995), with whom I worked in the first half of 1980s.

4. To make all the variant texts available for analysis. We are all aware of the so called "human factor;" we all for various reasons make mistakes. In order to mitigate errors we decided to place in the Corpus all forms of any single text. If one is unsatisfied with a standardized (normalized) version of a given text, one may have a look at an original version of it, for instance, in parallel with the standardized one. And if there are doubts that the original text has been read correctly, one has a chance to look at any page of the original manuscript or to listen to it (in cases where there is an audio record available). An example of such an audio-record can be found at http://corpora.iea.ras.ru/corpora/describe_text.php?id=28, where a short excerpt (25 minutes from a 2.5-hour recording) of the epos *Chylan-Toochii* recorded by me from Vladimir Tannagashev in 2003 is available, accompanied by a transcription completed by L. N. Arbačakova and myself in 2011.

The Corpus is an open resource with very soft terms of use. One may freely use short excerpts (normally, one to three sentences in length) from the texts of the Corpora as examples or illustrations, provided he/she gives due credit to the *Linguistic Corpora at IEA RAS* project and to the author and/or performer of the text used. If users want to reproduce longer fragments or whole texts in any form, contact should be established with the project leader or the person in charge of a specific corpus in order to negotiate the possibility and conditions of such usage.



Proper names in this text: Ай-Кан, Ак-Кожеге, Ак-Пилек, Алтын-Кылыш, Алтын-Шур, Кара-Кылыш, Кара-Салғын, Күн-Арыг, Күн-Кан, Күн-Каны, Чағыс-Чайачы, Чайачы, Челбиген.

View Options

- normalised text
- self-recording
- normalised text and original, with comments in parallel
- scanned image of original record (in new window)

comments are shown after a click on C to the rigth of the text

Ак-Пилек

	Normalised text	Original (self-recording)
1.	Амдығы төлдиң алында, пурунғу төлдиң соонда полча.	Амдыгы төльдиңь алында, пурунгу төльдиңь сонда полча.
2.	По чер пүдерде, чер-суг қабыжарда.	По чер пүдерде, чер-суг кабышарда
3.	Қалақпа чер пөлүжерде, қамышпа суг пөлүжерде полтур.	Калакпа чер пөльүшерде, камышпа суг пөльүшерде полтур.
4.	Чер ортазы черде қырық ашқымныг ақ тайға турча, ақ тайғаның төзүбе, толқуп келип, ақ талай ақ түшкен полтур по черде.	Чер ортасы черде, кырык ашкымныг ак тайга турча, ак тайганыңь төсүбе, толкуп келип ак талай ак түшкен полтур по черде.

Fig. 8. An original version of the epos *Ak-Pilek* given in parallel with the standardized (normalized) one, from http://corpora.iea.ras.ru/corpora/describe_text.php?id=19.

A short excerpt of the epos *Chylan-Toochii*, Vladimir Tannagashev, 2003. http://journal.oraltradition.org/issues/28ii/funk#myGallery-picture(14)

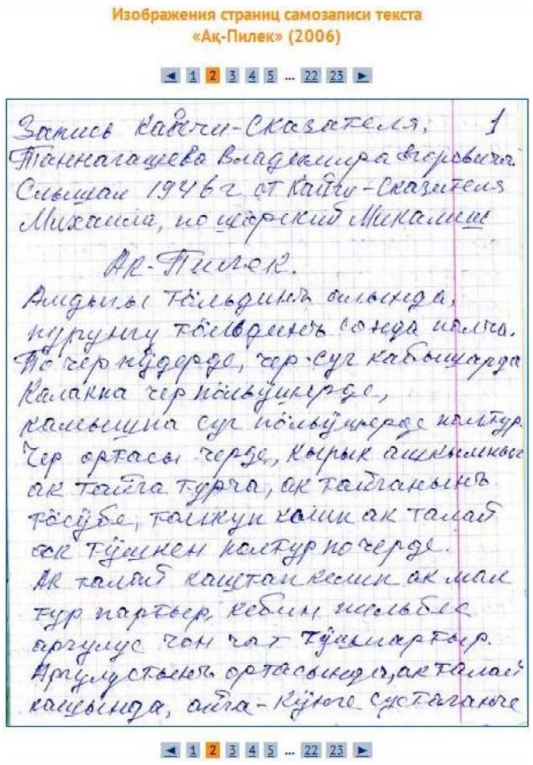


Fig. 9. A scanned page from Tannagashev's self-recording of the epos *Ak-Pilek*. See also http://corpora.iea.ras.ru/corpora/pages.php?id_text=19&page=2#image.

5. To make all texts in the Corpus searchable. In other words, the Corpus is not just a store and we are not just collecting texts in an electronic form, but we are making them available for research. For instance, we can choose texts collected in the Low-Mras region from Tannagashev's repertoire and recorded by him personally on paper. The result is a list of 21 epic tales. After that we can use the option "to compare. . ." and by clicking on this button we come to a page with four different options, where we can then create in one click a list of (standardized or original) word-forms a) common to all selected texts, b) present in the first but absent from other selected texts, c) present in at least one of the selected texts (thereby creating a dictionary of a given idiolect), and d) unique to the first selected text, that is, absent from every other text in the corpus. As soon as we get a list of all the word-forms we were seeking, we can save them as a file and work on the list later, whenever we want. We can therefore easily

- look up specific word-forms and their contexts;
- find word-forms directly preceding or succeeding the chosen one or find co-occurrences of given word-forms within a certain distance of each other within sentence;
- collect statistical data on the frequency of any word-form, comparing lists of word-forms from any number of texts in the same language;
- compare sentences from any two texts in order to identify recurring expressions of nearly any length. (One does not need to know all the possible expressions or *formulas* before finding them; the level of closeness can be set between 1 (=100%) and 0.2 (=20%) according to the researcher's preference.)

Fig. 10. A list of word-forms from the epos *Ak-Pilek* with examples, from http://corpora.iea.ras.ru/corpora/wordforms.php.

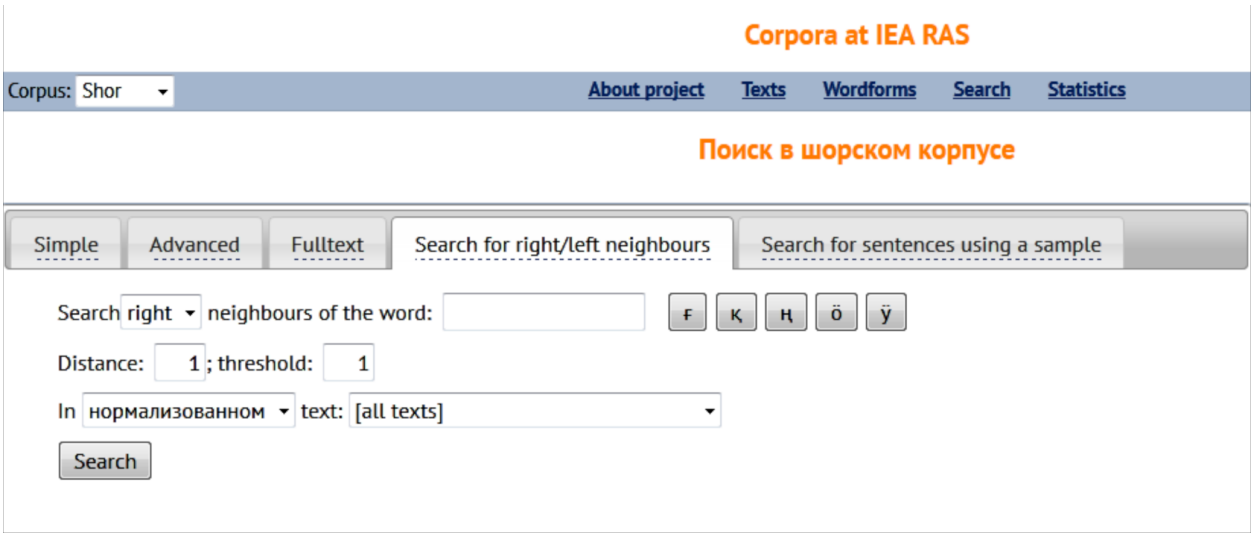


Fig. 11. Search options, from http://corpora.iea.ras.ru/corpora/search.php.

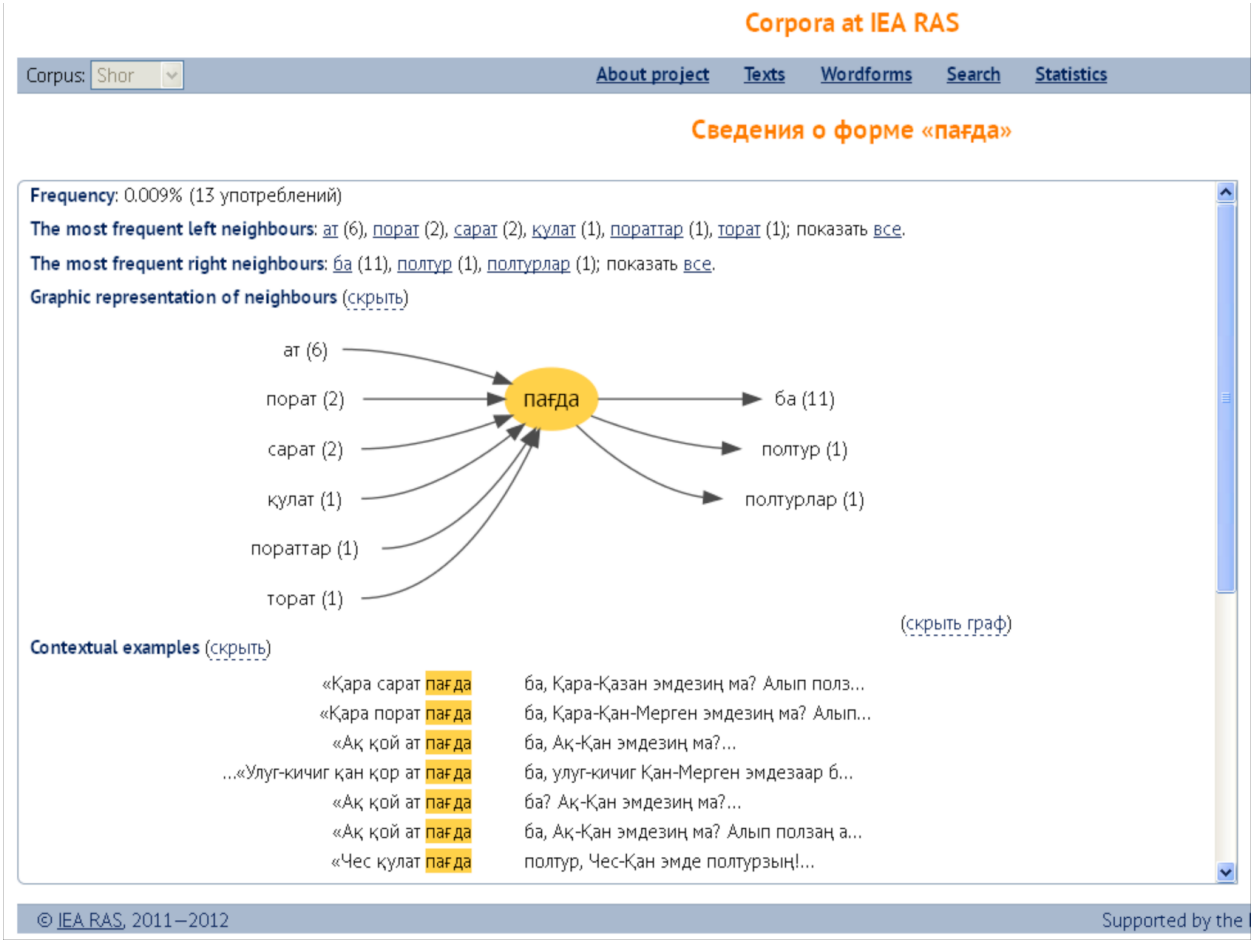
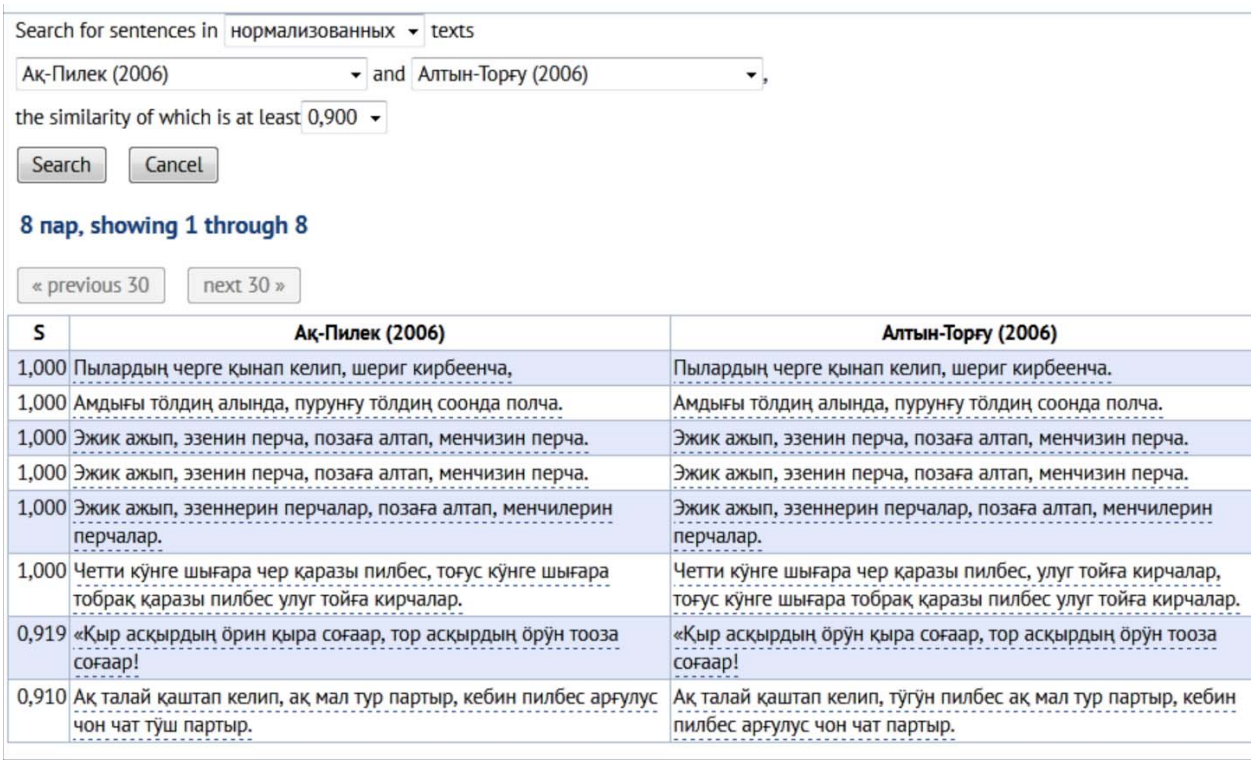


Fig. 12. Information about the word-form *pagda* (“on a leash”) (with graphic representation and contextual examples), from http://corpora.iea.ras.ru/corpora/describe_word.php?lang_code=cjs&wf_kind=normalised&word=%D0%BF%D0%B0%D2%93%D0%B4%D0%B0.



Search for sentences in нормализованных texts

Ақ-Пилек (2006) and Алтын-Торғу (2006),

the similarity of which is at least 0,900

Search Cancel

8 пар, showing 1 through 8

« previous 30 next 30 »

S	Ақ-Пилек (2006)	Алтын-Торғу (2006)
1,000	Пылардың черге қынап келип, шериг кирбеенча,	Пылардың черге қынап келип, шериг кирбеенча.
1,000	Амдығы төлдиң алында, пурунғу төлдиң соонда полча.	Амдығы төлдиң алында, пурунғу төлдиң соонда полча.
1,000	Эжик ажып, эзенин перча, позаға алтап, менчизин перча.	Эжик ажып, эзенин перча, позаға алтап, менчизин перча.
1,000	Эжик ажып, эзенин перча, позаға алтап, менчизин перча.	Эжик ажып, эзенин перча, позаға алтап, менчизин перча.
1,000	Эжик ажып, эзеннерин перчалар, позаға алтап, менчилерин перчалар.	Эжик ажып, эзеннерин перчалар, позаға алтап, менчилерин перчалар.
1,000	Четти күнге шығара чер қаразы пилбес, тоғус күнге шығара тобрақ қаразы пилбес улуг тойға кирчалар.	Четти күнге шығара чер қаразы пилбес, улуг тойға кирчалар, тоғус күнге шығара тобрақ қаразы пилбес улуг тойға кирчалар.
0,919	«Қыр асқырдың өрин қыра соғаар, тор асқырдың өрўн тооза соғаар!	«Қыр асқырдың өрўн қыра соғаар, тор асқырдың өрўн тооза соғаар!
0,910	Ақ талай қаштап келип, ақ мал тур партыр, кебин пилбес арғулус чон чат түш партыр.	Ақ талай қаштап келип, түгүн пилбес ақ мал тур партыр, кебин пилбес арғулус чон чат партыр.

Fig. 13. Recurring expressions in the eposes *Ak-Pilek* and *Altyn-Torgu* (with the level of similarity of 0.900 and higher), from http://corpora.iea.ras.ru/corpora/compare_texts.php.

The standards we initially set for the Corpus can and must be improved. We began the project with the aim of presenting only folklore texts from three ethnic groups (the Shors, the Teleuts, and the Evenki). But now, thanks to the constantly growing interest of our colleagues, we have expanded the Corpus's borders: it already contains a good number of Nenets texts and some Bashkir texts. What is more, we have changed the main goal of the Corpus to include samples of professional literature, newspapers, and religious as well as other types of texts.

It is hard to predict what the future holds in store for our Corpus. But even now it is quite obvious that the Corpus—realized initially as a simple database—offers a unique possibility to analyze folklore texts in many different ways, making it of especial value for linguists, folklorists, and cultural anthropologists alike.

Moscow State University,
National Research Tomsk State University

References

Arbačakova 2001 — L. N. Arbačakova. *Tekstologija shorskogo geroicheskogo eposa (na primere materialov N. P. Dyrenkovoj i A. I. Chudojakova*) (*Textology of the Shor Heroic Epics* [by the example of N. P. Dyrenkova's and A. I. Čudoyakov's materials]). Novosibirsk: Nauka.

Baskakov 1981 — N. A. Baskakov. *Altajskaja sem'ja jazykov i ee izuchenie* (*Altai Language Family and its Study*). Moscow: Nauka.

Čudoyakov 1995 — A. I. Čudoyakov. *Etjudy shorskogo eposa* (*Sketches of the Shor Epics*). Kemerovo: Kemerovo Publishing House.

Dyrenkova 1940 — N. P. Dyrenkova, ed. and trans. *Shorskij fol'klor* (*Shor Folklore*). Moscow and Leningrad: Academy of Sciences of the USSR.

Esipova and Arbačakova 2006 — A. Esipova and L. Arbačakova. "Archaic Vocabulary in Shor Heroic Epics." In *Exploring the Eastern Frontiers of Turkic*. Ed. by M. Erdal and I. Nevskaya. Wiesbaden: Harrassowitz. pp. 19-40.

Funk 1999 — D. A. Funk. "Zametki na poljax shorsko-russkogo slovarja" ("Notes on the Margins of a 'Shor-Russian Dictionary'"). In *People of the Russian North and Siberia*. Ed. by Z. P. Sokolova and D. A. Funk. Moscow: Staryj Sad. pp. 141-67.

Funk 2003 — _____. "Molochno-belye koni v epose tajezhnyx oxotnokov, rybolovov I sobiratelej" ("Milk-White Horses in the Epic Stories of Taiga Hunters, Fish Men and Gatherers"). *Etnograficheskoye obozreniye*, 3:53-60.

Funk 2005 — _____. *Miry shamanov i skazitelej* (*Worlds of Shamans and Storytellers*). Moscow: Nauka.

Funk 2006 — _____. "Anthroponymic Patterns in Traditional Social Culture and in Epic Texts." In *Exploring the Eastern Frontiers of Turkic*. Ed. by M. Erdal and I. Nevskaya. Wiesbaden: Harrassowitz. pp. 41-57.

Funk 2010-13 — _____, ed. and trans. *Shorskij geroicheskij epos* (*Heroic Epics of the Shors*). 4 vols. Moscow: Institute of Ethnology and Anthropology, Russian Academy of Sciences and Kemerovo: Primula.

Funk and Tomilov 2006 — D. A. Funk and N. A. Tomilov, eds. *Tjurkskie narody Sibiri* (*Turkic Peoples of Siberia*). Moscow: Nauka.

Radloff 1866 — W. Radloff. *Obrazcy narodnoj literatury tjurkskich plemen: živuščich v Južnoj Sibiri i Dzungarskoj stepi* (*Patterns of Folk Literature of the Turkic Tribes Living in Southern Siberia and the Dzungarsky Steppe*. Part 1. St. Petersburg: Tipografija Imperatorskoj Akademii Nauk.

This page is intentionally left blank.