New Kazakh parallel text corpora with on-line access

Zhandos Zhumanov[1], Aigerim Madiyeva[2] and Diana Rakhimova[3]

al-Farabi Kazakh National University, Laboratory of Intelligent Information Systems, Almaty, Kazakhstan

[1]z.zhake@gmail.com, [2]rockinfuture_7@mail.ru,
[3]di.diva@mail.ru

Abstract. This paper presents a new parallel resource – text corpora – for Kazakh language with on-line access. We describe 3 different approaches to collecting parallel text and how much data we managed to collect using them, parallel Kazakh-English text corpora collected from various sources and aligned on sentence level, and web accessible corpus management system that was set up using open source tools – corpus manager Mantee and web GUI KonText. As a result of our work we present working web-accessible corpus management system to work with collected corpora.

Keywords: parallel text corpora, Kazakh language, corpus management system

1 Introduction

Linguistic text corpora are large collections of text used for different language studies. They are used in linguistics and other fields that deal with language studies as an object of study or as a resource. Text corpora are needed for almost any language study since they are basically a representation of the language itself. In computer linguistics text corpora are used for various parsing, machine translation, speech recognition, etc. As shown in [1] text corpora can be classified by many categories:

- Size: small (to 1 million words); medium (from 1 to 10 million words); large (more than 10 million words).
- Number of Text Languages: monolingual, bilingual, multilingual.
- Language of Texts: English, German, Ukrainian, etc.
- Mode: spoken; written; mixed.
- Nature of Data: general, specialized (dialect, idiolect, sociolect, etc.).
- Nature of Application: research, illustrative, learner, translation, aligned comparable, parallel, reference.
- Dynamism: dynamic (monitor), static.
- Temporal characteristic: diachronic, synchronic.
- Authorship: one author, two and more.
- Annotation: unannotated, annotated (morphologically, semantically, syntactically, prosodically, etc.).

- Access: free, commercial, closed.

There are quite a lot of text corpora for different languages. Among them are:

- American National Corpus;
- British National Corpus;
- Brown Corpus;
- Russian National Corpus;
- Europarl Corpus;
- EUR-Lex corpus.

There are number of text corpora for the Kazakh language. But all of them are monolingual:

- Almaty Corpus of Kazakh;
- Kazakh text corpora on Sketch Engine;
- Open-Source-Kazakh-Corpus;
- Kaz Corpus [2].

At the moment there is not a lot of parallel data that involves Kazakh language. Also the data is presented in raw format, usually it is a plain text files with one sentence on each line. Files in different languages are aligned on sentence level.

Parallel text corpora are very useful for comparative studies of all kinds. But they are also much more difficult to gather. Since so little ready to use parallel text corpora exist for Kazakh language there is a clear need to collect them somehow. That is the first task of this research.

When collected, the text corpora have to be accessed and worked with. Raw plain text formats are good for computers, but not for humans. Some system has to be put in place to facilitate the text corpora. Setting up such system is the second task of this research.

Following sections are dedicated to more detailed description of: section 2 – the process and approaches of collecting parallel text corpora we have used; section 3 – analysis of text corpora we have gathered; section 4 – description of corpus management setup; section 5 – conclusion and future work.

2 Process of parallel text corpora collection

In order to collect parallel text corpora we used 3 different approaches:

1. finding all significant ready to use aligned parallel texts;
2. using bitextor tool for crawling websites that contain same texts in several languages and aligning them;
3. using scrips for crawling texts from websites that contain same texts in several languages and using InterText tool with integrated hunalign tool for aligning them.

Approaches 2 and 3 seem to be similar to each other, but they have produced different amount of results which is described below.

There are not many places to find ready to use parallel text corpora that have Kazakh as one of the languages. In fact, there is one such place - the OPUS project [3]. There were some parallel Kazakh-English texts collected from Tatoeba and OpenSubtitles. That gave us 4480 aligned sentences.

Another ready to use resource is the Bible. It has been repeatedly translated into many languages. The Kazakh is also among them. There were several translations prepared by several organizations. The most resent one is called "New World Translation of the Christian Greek Scriptures" published on different media by Jehovah's Witnesses. Despite the nature of the organization it is turned out to be a great parallel resource since the text of the book has strictly numbered chapters and verses across all translations. That provided us Kazakh-English parallel text of 32358 sentences.

Bitextor is a free open source application for collection of translation memories from multilingual websites. [4] The application downloads all HTML files from a website, then pre-processes them into a consistent format and applies a set of heuristics to select the file pairs that contain the same text in two different languages (bitexts). Using LibTagAligner library translation memories in TMX format are created from these parallel texts. The library uses HTML-tags and length of the text segments for alignment. After cleaning the resulting translation memory from TMX format tags, we receive a parallel corpus with sentences in different languages aligned with each other.

We have run bitextor for following websites: http://www.kaznu.kz, http://www.bolashak.gov.kz, http://www.enu.kz, http://www.kazpost.kz, http://www.archeolog.kz, http://e-history.kz, http://inform.kz, http://egov.kz, http://primeminister.kz, http://tengrinews.kz and etc. (fig. 1).

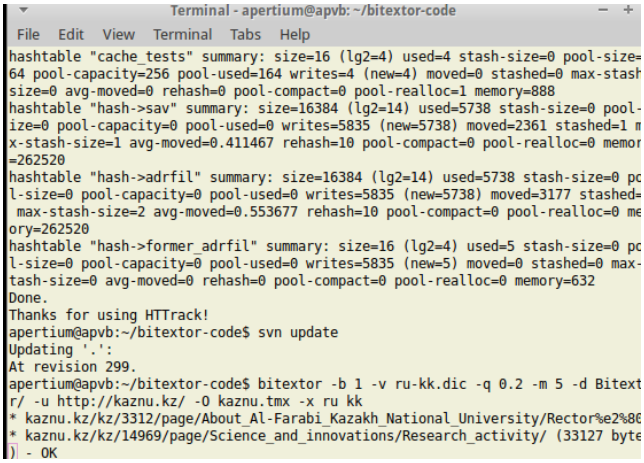

Fig. 1. An example of running bitextor for www.kaznu.kz

As a result of bitextor's work from each site we obtained *.tmx file with the following format (fig. 2):

Fig. 2. A format of obtained parallel corpus for Kazakh-Russian language pair

In this format, tag <tu> includes a pair of aligned segments (in this case - sentences); tag <tuv> - separate sentences in two languages; tag <prop> - HTML file addresses from which these sentences have been extracted; tag <seg> - sentences themselves. In such *.tmx file sentence in one language corresponds to the sentence in another language. It should be noted that comparison quality depends on the website. Thus, we receive a file with parallel texts.

During cleaning of TMX files recurring segments, erroneous and meaningless sentence pairs were deleted. After removal of tags, we received Kazakh-English parallel corpus with 5 925 sentences.

The third approach is partially automated but also involves manual checking of the results. It consists of following stages:

- crawling parallel texts on the internet;
- cleaning and formatting of gathered texts;
- sentence splitting;
- sentence alignment;
- manual checking.

All stages except the last one can be automated. But the quality of the parallel text will affect quality of the tasks to be solved with them. So in our opinion human involvement is mandatory.

As a source for parallel texts we used web-sites http://www.akorda.kz/ and https://www.ted.com/. Texts from the first one were collected using scripts links.pl and extract_text.pl that are available in Apertium project's repository using the following link: https://sourceforge.net/p/apertium/svn/59905/tree/languages/apertium-kaz/texts/akorda/. Texts from the second site were collected manually.

After cleaning, formatting and sentence splitting we had two lists of sentences in two languages that were translations of each other but the sentences themselves were not aligned due to various translation reasons. To align them we used hunalign tool

[5]. Hunalign has remarkably high quality: we got 6-8% of incorrectly aligned sentences out of unaligned lists mentioned above. But low percentage still meant that we had 2000-3000 alignment mistakes. It is quite many and that is why manual checking was due. Parallel text alignment editor called InterText was used for that (fig. 3). [6]

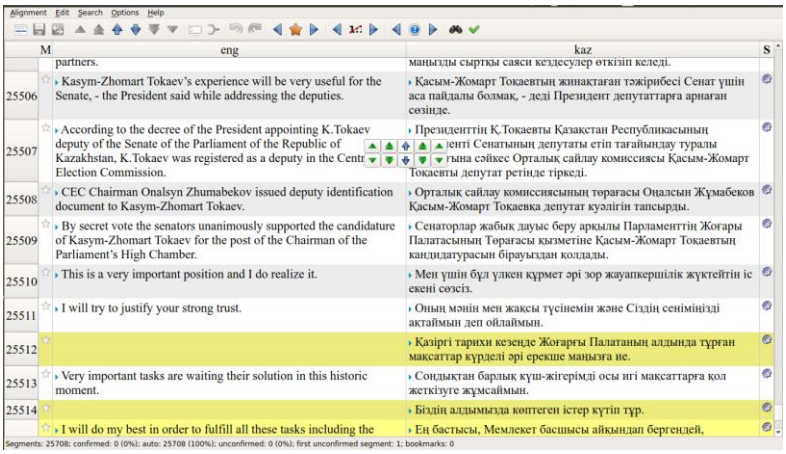

Fig. 3. InterText parallel text alignment editor

Approach described above resulted in two text corpora:

1. Akorda - 24 148 aligned sentences.
2. TED - 6 120 aligned sentences.

It seems to be logical that we should have tried to use bitextor on the Akorda and TED sites. And so we did. But for some unknown reasons bitextor did not produce any results or produced very few (under 100) aligned sentences when we tried to apply it with different settings.

All the raw text corpora described in this section are available on: https://drive.google.com/drive/folders/0B3f-xwS1hRdDM2VpZXRVblRRUmM.

3 Analysis of collected text corpora

Information about all the text corpora that we have gathered is provided in the table below.

Table 1. Description of gathered text corpora

#	Method	Corpus	# of sentence pairs	# of words kaz	# of words eng
1	Ready to use	OPUS	4 480	19 892	27 839
2	Ready to use	New World Bible	32 358	548 258	824 398
3	Bitextor	Lab IIS	5 925	112 658	157 313
4	Semi-manual	Akorda	24 148	341 154	456 689

5 Semi-manual	TED	6 120	54 965	79 320
	TOTAL:	73 031	1 076 927	1 545 559

According to the classification shown earlier we gathered medium sized, bilingual, Kazakh-English, written, general text, aligned, parallel, static, many author, unannotated, free text corpora.

4 Setting up a corpus management system

Our second task – setting up a system to work with parallel text corpora – has been achieved with the help of an open-source tool called Manatee. [7] It is employed as the main corpus management tool for several large text corpora, including the Czech National Corpus, and we plan to use this system as the basis to maintain the parallel text corpora for Kazakh-English and Kazakh-Russian language pairs. Manatee is able to deal with extremely large text corpora and is able to provide a platform for computing a wide range of lexical statistics. It has such features as text preparation, concordancing, meta-data management, tokenization, efficient corpus storage, corpus annotation, computation of statistics and it is language and tag- set/annotation independent. Moreover, Manatee also functions as a corpus management server, which satisfies the condition of the web-accessibility of the corpus manager. To access its features we will use the GUI called KonText. It is a fully featured corpus query interface for the Manatee corpus search engine. It started as an extension of the Bonito 2.68 web interface but now is gradually becoming more independent. It is maintained by the Institute of the Czech National Corpus and the source code of the project is available at the URL: https://github.com/czcorpus/kontext/. All the key features of the Bonito 2.98.3, primarily a support for parallel text corpora, is present in the current version of the interface.

At first we've tried to implement the system that combines Manatee and Bonito into a corpus management tool called NoSketch Engine. We installed and locally hosted it during the testing period but since KonText has a lot of new features, more enhanced user interface and improved code documentation, we decided that it will be more reasonable to switch from Bonito to KonText. KonText comes with default plug-ins (located in lib/plugins directory), which provide a complete, working set of replaceable components needed to run it with all the features enabled but for the time being we have made it work with only the basic functionality such as search, concordances and frequency analysis, our main focus being the ability to manage parallel texts.

For now the system is hosted on Google App Engine virtual machine as the service provides flexible and inexpensive platform to experiment with different setups. It can be reached at http://104.197.218.108/corpora/corplist. Figure 4 shows a page from the system with list of all available corpora.

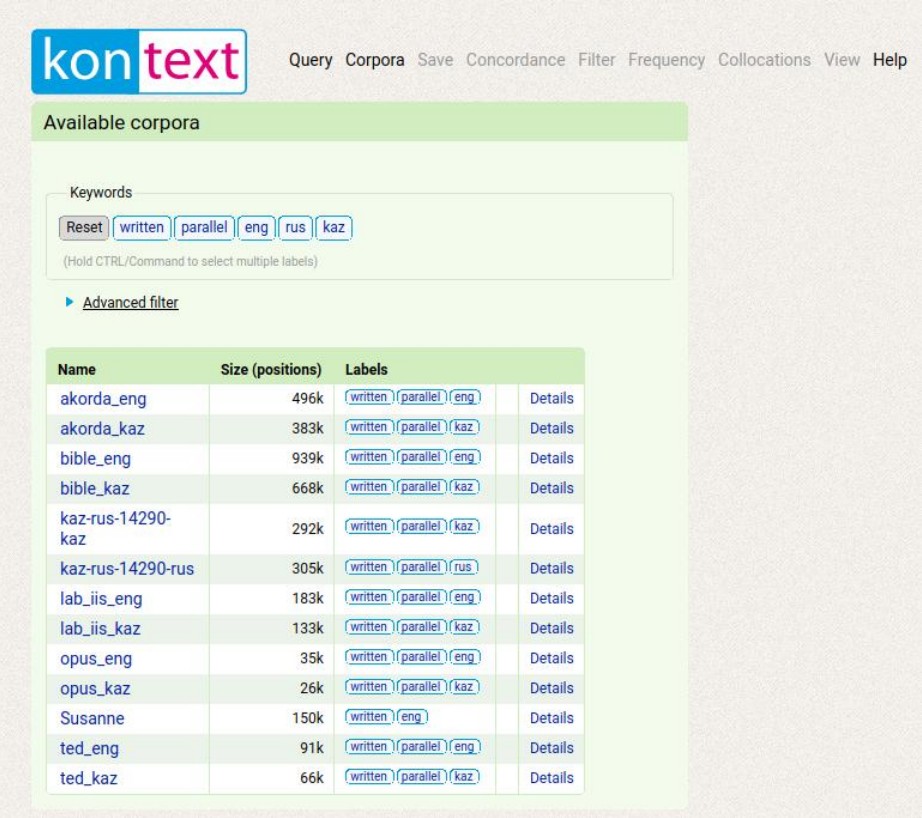

Fig. 4. List of corpora available via corpus management system

5 Conclusion

We have presented new parallel text corpora for Kazakh language with on-line access. Using 3 different approaches to collecting parallel data we have gathered medium sized, bilingual, Kazakh-English, written, general text, aligned, parallel, static, many author, unannotated, free text corpora. The corpora are available in raw text format along with web accessible corpus management system that is based on corpus manager Mantee and web GUI KonText.

For future we plan to continue work on the corpus and corpus manager. One direction of our efforts will cover collecting and possibly creating more parallel data including Kazakh and other languages. Another direction will cover implementation of all the other functionalities available in KonText and/or extend it even further at the next stages of our work.

References

1. Sereda, Iryna. "Approaches to corpora classification in modern Corpus Linguistics." (2012).
2. Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov, B., Sabyrgaliyev, I., & Sharafudinov, A. (2013, October). Assembling the Kazakh Language Corpus. In EMNLP (pp. 1022-1031).
3. Tiedemann, Jörg, and Lars Nygaard. "OPUS-an open source parallel corpus." Proceedings of the 13th Nordic Conference on Computational Linguistics (NODALIDA). University of Iceland, Reykjavik. 2003.
4. Esplá-Gomis, Miquel, and Mikel L. Forcada. "Bitextor, a free/open-source software to harvest translation memories from multilingual websites." Proceedings of MT Summit XII, Ottawa, Canada. Association for Machine Translation in the Americas (2009).
5. D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages In Proceedings of the RANLP 2005, pages 590-596.
6. Vondřička, Pavel (2014): "Aligning parallel texts with InterText" In: Calzolari, N. et al. (ed.): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA). p. 1875-1879.
7. Rychlý, P. (2007, December). Manatee/bonito-a modular corpus manager. In 1st Workshop on Recent Advances in Slavonic Natural Language Processing (pp. 65-70).