The Aligned Kazakh-Russian Parallel Corpus Focused on the Criminal Theme

Nina Khairova[1][0000-0002-9826-0286], Anastasiia Kolesnyk[1][0000-0001-5817-0844], Orken Mamyrbayev[2][0000-0001-8318-3794], Kuralay Mukhsina[3][0000-0002-8627-1949]

[1]National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., 61002, Kharkiv, Ukraine
[2]Institute of Information and Computational Technologies, 125, Pushkin str., 050010, Almaty, Republic of Kazakhstan
[3]Al-Farabi Kazakh National University, 71 al-Farabi Ave., Almaty, Republic of Kazakhstan,
khairova@kpi.kharkov.ua, kolesniknastya20@gmail.com,
{morkenj, kuka_ai}@mail.ru

Abstract. Nowadays, the development of high-quality parallel aligned text corpora is one of the most relevant and advanced directions of modern linguistics. Special emphasis is placed in creating parallel multilingual corpora for low resourced languages, such as the Kazakh language. In the study, we explored texts from four Kazakh bilingual news websites and created the parallel Kazakh-Russian corpus of texts that focus on the criminal subject at their base. In order to align the corpus, we used lexical compliances set and the values of POS-tagging of both languages. 60% of our corpus sentences are automatically aligned correctly. Finally, we analyzed the factors affecting the percentage of errors.

Keywords: criminal subject, news websites, POS-tagging, Kazakh-Russian parallel corpus, alignment, lexical compliances.

1 Introduction

As of today, linguistic resources are not only a part of any linguistics study but an important base for designing any NLP applications. Such resources typically include dictionaries, thesauri, linguistics ontologies, monolingual and multilingual corpora. In order to create these linguistics resources lexicographic researches, analysis of the lexical structure of languages, exploring the text characteristics and similar studies are being conducted.

Design and creation, development and use of high-quality text corpora are one of the most relevant and advanced directions of modern linguistics [1]. Such processed and systematized by means of concordancer corpus allows storing a large amount of text information necessary for the statistical analysis of the linguistic phenomena and diachronic change in spoken and written languages.

There are a lot of types of corpora. There are specialized corpora (genre, time, place), general corpora, multilingual corpora, learner corpora, historical or diachronic

corpora, monitor corpora and multilingual corpora. Multilingual corpora, in turn, are divided into comparative (comparable corpus) and parallel or the corpus of the translations (translation corpus).

In our opinion, parallel text corpora are particulary important in studying language and features of the translation, various parsing, tasks of speech recognition, etc. For instance, in the tasks of foreign language training, such corpora allow finding possible equivalents of the analyzed lexicon, tracking its values and functions in some contexts.

Furthermore, the concept of the parallel corpus is an integral part of the broader and more difficult concept, such as – machine translation. It is known that machine translation is still the unresolved task of computational linguistics, despite the rapid growth of the various program and empirical resources. In some times the quality of machine translation also depends on the amount of parallel sentences used in training.

For the last decade in the world there was created the set of bilingual and multilingual corpora, among which, in our view, the most exciting are: EUROPARL-20.000.000 word usage, the open corpus of European Parliament in 11 languages,) (https://www.isi.edu/~koehn/publications/europarl/); CHEMNITZ GERMAN-ENGLISH TRANSLATION CORPUS – 1.000.000 word usage (http://www.tu-chemnitz.de/phil/InternetGrammar); KACENKA (Korpus anglicko-cesky; Czech) - 3.000.000 word usage (http://www.phil.muni.cz/angl/kacenka/kachna.html); OPUS - (5 languages) (https://aclanthology.info/papers/L04-1174/l04-1174); English-French Canadian Hansard [2].

The most prominent and the greatest Kazakh language corpora are: Almaty Corpus of Kazakh (http://web-corpora.net/KazakhCorpus/search/), containing more than 40 million word usage, 86% of word usage have grammatical analysis; Kazakh text corpora on Sketch Engine [3]; Open-Source-Kazakh-Corpus, created with the use of the Wikipedia dump tool and including a collection of 20 million words (600 thousand of them are unique) [4]; Kazakh Language Corpus (KLC) [5].

At the same time, despite the existence of a large number of parallel multilingual corpora, for low resourced languages, such as the Kazakh language, the task of parallel corpora creating is vital. The task becomes more complex when we say about the development of parallel corpora for not similar languages, the languages from different families. For instance, one language belongs to the Turkic language family and the other belongs to the Indo-European language family, as Kazakh and Russian.

In our study, we explored texts in two languages (Russian and Kazakh) from Kazakh bilingual news websites and created the parallel Kazakh-Russian corpus at the base of these sites' texts. Moreover, texts of our parallel corpus do not belong to fiction or another broad theme; they focus on the criminal subject that makes them limited-field. Therefore, we were able to apply the dictionary method to align the corpus. In addition, to improve the quality of sentence alignment we have made POS-tagging of the texts in both languages and then exploit the labelling. Finally, we calculate the percentage of correct aligned sentences and analyzed the factors affecting the percentage of error.

2 Related Work

Parallel corpora contain the text of the original and its translation into some other language. Additionally, these two texts are not just opposed each other, they have to be aligned: particular fragments of the original text have to coincide with the corresponding fragments of the translation. We can say that a parallel corpus is only useful when it is aligned.

In most studies, two levels of alignment are explicitly or implicitly distinguish: sentence alignment and lexical alignment. Generally, the task of automatic comparison of sentences or words in one text to their equivalents in translation is very labor intensive as this consistency between words or sentences is sometimes not "one to one". For instance, a few paragraphs in source language can correspond to one paragraph in the target language, in translation some words can be deleted or replaced with very distant synonyms or fixed phrases which can be absolutely various in different languages, etc.

We can classify sentence alignment methods into 3 three categories. Methods of the first category are based on the use of lengths of sentences and paragraphs [2]. This approach uses a hypothesis that the length of the sentence in the original and in translation approximately match.

The second group of methods uses lexical information, which can be received from the corpus [6], [7]. Unfortunately, these methods are applied extremely rarely, which is due to inaccessibility of bilingual dictionaries and difficulties of the automatic morphological analysis to mutual identification of words in dictionaries. To date, most of applications based on this group of methods exploit only texts of specialized subjects, for example, texts of parliaments and legal texts. The use of dictionary methods for literary texts is rare because even in a similar genre there is a high percentage of the ambiguity of vocabulary in compared sentences.

The third group of texts alignment algorithms in parallel corpora is based on the POS-tags which are contained in an annotated corpus or use spelling similarity [8].

However, the use of any method of these groups has a number of inaccuracies and weaknesses. Accordingly, nowadays interest in the development of systems which apply the combination of all approaches constantly grows. For instance, Varga et al. [9] described a hybrid method of parallel text alignment. They used the alignment technique that combines the length-based method with some kind of translation-based similarity. The basis of research was formed by Hungarian, Romanian, and Slovenian languages.

Rico Sennrich and Martin Volk [10] in their study showed that sentence alignment can be achieved without the use of language-specific resources other than the to-be-aligned parallel text. They used a length-based sentence alignment algorithm and train an SMT system on the to-be-aligned text. Such a system is used to translate the source side of the parallel training corpus and then it bases its sentence alignment on this translation. In their study, they proved that the iterative sentence alignment approach leads to the best results after just two iterations.

Another approach to sentence alignment is described in the article [11]. In this paper, authors proposed their own fast and robust sentence alignment algorithm − Fast-

Champollion, which employs a combination of both length-based and lexicon-based algorithm. The method is called "fast" because it optimized the process of splitting the input bilingual texts into small fragments for alignment. This method needs a dictionary for aligning sentences, but its precision and recall will drop as the size of the dictionary decreases.

Vondricka [12] made the review of a special application InterText for alignment of parallel corpora that is based on some hybrid alignment methods. This resource exists in two forms: InterText Server (server based on the text management system with web-based editor interface) and InterText editor (personal desktop application). Both are open-source software. The same application was used for the creation of Kazakh-English text corpora in the study by Zhumanov et al. [13]. Furthermore, authors exploited Bitextor, and hunalign tools in order to crawl websites that contain the same texts in several languages and aligned them. The article by Rakhimova et al.[14] is devoted to the meeting similar challenges. They considered the principles of use of such application as Bitextor, which generates translation memories using multilingual websites as a corpus source. It downloads an entire website and applies a set of heuristics (based mainly on HTML tag structure and text block length) to find bitexts.

Authors of the article [15] aligned their dataset at the sentence level, it means they used strong punctuation for the text segmentation. But such an approach de-mands verification of text correctness and proximity of languages. So the manual control is required. As a result, all medical and all literary texts in the Polish/Ukrainian pair has been aligned and verified, while only part of French and English texts is still being operated. At the same time, Finnish and Russian versions of the Aranea corpora and the newspaper subcorpus of the Russian national corpus and a corpus of the Finnish national library are aligned sufficiently well [16].

The creation of aligned parallel corpora faces the additional challenge of the search of textual resources for a parallel corpus. Nowadays there are a lot of researches relating to obtaining parallel sentences from non-parallel or comparable data. For example, such linguistic data can be obtained from Wikipedia. This is an extremely valuable resource for extracting parallel sentences, as the document alignment is already provided and Wikipedia articles on the same topic can be in different languages. In addition, they are connected via "interwiki" links, which are annotated by users [17]. However, Wikipedia has not been thoroughly explored yet [18].

Consequently, we can make a conclusion that the problem of parallel corpora alignment has still not been solved up to the end and the universal method has not been found. Furthermore, to date, most studies consider that the choice of the alignment method depends on the researched language pair, thematic focus of texts and types of documents represented in the corpus [19].

3 Dataset preparation

The parallel corpus creation task includes several parts. The first and one of the most important tasks is to collect text material for such corpora. Despite the fact that the Internet contains a large number of websites which are bilingual and multilingual, the choice of needed bilingual resources constitutes an important part of the parallel cor-

pus elaboration. This task becomes more complicated by the fact that we process such different languages as Kazakh and Russian. Additionally, it is necessary to use specialized tools and various techniques not only for language processing but also for collecting necessary material for corpora.

We suggest that parsing of the websites is the best way to automate the process of collecting and saving information. For our study, we have developed our own special software for automatic parsing of the websites, which allow parsing websites that can be similar in design, content and structure.

Four bilingual websites zakon.kz caravan.kz, lenta.kz, nur.kz were chosen for the developed parser. The selected websites represent well-known and reliable news websites of Kazakhstan that are the main news sources on the criminal subject. They contain a large number of articles according to the criminal information, for example, different offences such as robberies, car thefts, murders, car incidents and others, which is one of the goals of our study. Additionally, the websites can switch text information between two languages: Russian and Kazakh.

As a result of a program runtime, we have received the general set of 3000 texts in two languages: Russian and Kazakh. From them, we have selected manually the test set for the creation of the aligned parallel corpus of the Russian-Kazakh texts on criminal information. The corpus size is more than 50410 words, about 24800 of them in Russian and about 25600 words in Kazakh.

On the next step, we determined the structure of the corpus organization. Nowadays such structure can be very diverse, depending on the pragmatical purposes of its creators or users:

- in the form of the traditional text with reference to the translations,
- in a tabular "mirror" form that is more convenient for perception and comparison,
- in the form of the database.

For our study, we chose the database structure as it is the most convenient way for storage of a large amount of data with a possibility of its further increasing.

All news articles are stored in the table of the database which includes their ID, title, the address of the website and the text of an article.

At the following stage, we carried out POS-tagging of the corpus. For a Russian corpus labelling, we chose the pymorphy2 (https://nlpub.ru/Pymorphy) Python packet which is specially developed for morphological analysis of Russian and Ukrainian texts. The libraries of the packet use the OpenCorpora (https://www.pydoc.io/pypi/gensim-3.2.0/autoapi/corpora/dictionary/index.html) dictionary and make hypothetical conclusions for non-recognized words.

In turn, the complexity, structural and typological characteristic of Kazakh marking is connected with the fact that it belongs to agglutinating languages. Structure of this language is rather difficult and unusual, since your native language is inflectional. The agglutinative formation is opposite inflectional where every formant has several inseparable meanings at once (for example, a case, gender, number, etc.). In this reason, we make POS-tagging of Kazakh texts via the regular expression tagger based on RegexpTagger class of nltk Python (https://www.nltk.org/) package. For example, we

can identify some types of nouns in Kazakh texts via the following list of regular expressions:

```
patterns=[(r'.*бен$','NN'), (‘r.* пенен $','NN'), (‘r.*
басшылық $','NN'), (r'.* іпқону $','NN'), (r'.* тармен
$','NN'), (r'.* герлермен $','NN'), (r'.* здар $','NN')]
```

Additionally, to increase recall and precision of our POS-tagging of Kazakh texts we combine regular expressions with the system that includes seven rules. For instance, "If a word followed by words from the special list — the word is marked as Verb".

4 The automatic alignment of the corpus

At the first step of the automatic alignment of our corpus, we were guided by punctuation symbols, capital letters and paragraphs. At the next step, it is possible to select two basic approaches to sentence alignment. The first approach that provides significantly higher productivity is based on sentence length. In the second, more resource-intensive approach, the lexical compliances set in by a word alignment method. In our research, the first approach will not yield exact and objective results as the Kazakh language is agglutinating. It means that the form of a word is formed by addition affixes as well as auxiliary additional words carrying semantic and morphological information. In this reason, the use of alignment on the length of sentences or paragraphs of inflectional and the agglutinating languages is not an effective method.

Upon detailed studying of this area, we revealed that for the languages belonging to different language groups and further for specialized, thematic texts it is the best of all to apply the dictionary method of alignment. On the basis of this conclusion, we exploit the lexical compliances set and the values of POS-tagging obtained in previous stages of preparation. The main reason why we were not able to use the first easier approach to sentences alignment is a huge difference between syntax and semantics of Kazakh and Russian languages.

As a lexical compliances set we use our own Kazakh-Russian dictionary, which is based on the English-Kazakh-Russian dictionary that contains about 50 000 entries. Figures 2 shows the fragment of the English-Kazakh-Russian dictionary, which we use as a background one. The dictionary contains about 50 000 entries. Figures 3 shows the fragment of Kazakh-Russian dictionary, which we apply to align parallel Kazakh - Russian corpus.

```
native###_туасынан, ·жаратылысынан, ·тумысынан###_прирождённый·
(прирожденный)¶
native###_табиғи, ·жаратынды###_природный¶
native###_тұнық, ·ұқыпты, ·кірсіз, ·пәк, ·мөлдір, ·кіршіксіз, ·бейкүнә, ·
әйнектей, ·ақ, ·мұнтаздай, ·таза, ·шайдай·ашық, ·мөлдір·бұлақ, ·дақсыз, ·саф, ·
ашық, ·адал, ·шын, ·нағыз, ·айна·дай, ·бәкізе, ·ғилман, ·жазықсыз###_чистый¶
native###_жергілікті·адам, ·туып-өскен, ·туған###_уроженец¶
native###_туған, ·бір·туған, ·тіған-туысқандар, ·туыстық, ·туып·өскен, ·
қарағым, ·шырағым###_родной¶
native·of·a·country·(·aborigine·)###_абориген###_абориген¶
nativity·(the·Nativity)###_рождество###_рождество¶
nativity###_дүниеге·келу, ·жаралу###_рождение¶
north·atlantic·treaty·organization·(NATO)###_°###_североатлантический·
союз·(НАТО)¶
North·Atlantic·Treaty·Organization·(NATO)###_°###_Североатлантический·
союз·(НАТО)¶
nato·(NATO;·North·Atlantic·Treaty·Organization)###_°###_нато·(НАТО;·
```

Fig. 1. The fragment of the used basic English-Kazakh-Russian dictionary.

kz	ru
қылмыстық	криминал
мәйіті	трупы
мәліметтерге	данные
министрлігімен	с министерством
Оңтүстік	юг
орындарын	мест
өзара	взаимное
өзінің	его собственный
пиғылды	недобросовестный
Рудныйда	в Рудном
сату	продажа
сәттері	моменты

Fig. 2. The fragment of the texts store database in two languages.

To improve the sentence alignment, apart from the dictionary method, we use knowledge about the POS-tagging of the words in sentences. Such an approach will allow improving results of the dictionary method as the Kazakh words have several variants of the translations. Thanks to the correct marking of the words in the texts it is possible to improve the best translation equivalent.

5 Experiments and results

Our parallel Kazakh-Russian corpus contains texts from certain Kazakh news sites for the period 2018 – 2019. The corpus includes more than 50410 words, about 24800 of them in Russian and about 25600 words in Kazakh.

In order to assess the correctness of aligned sentences, we leverage experts' opinion, which are native speakers of the Kazakh language as well as the Russian language.

A well-designed special application allows the experts to choose the text in any (Russian or Kazakh) language and automatically load the parallel file of text. When working with a corpus, the expert may mark texts, save them with marking and align them manually. Figure 4 shows the user interface of our special universal application for working with aligned parallel corpora.

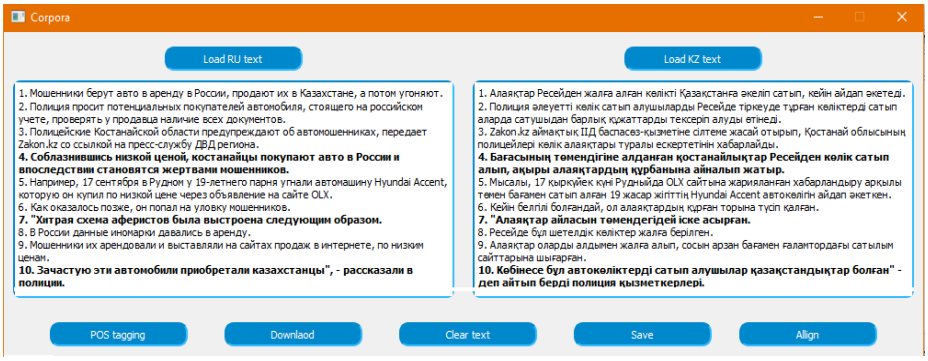

Fig.4. The user interface of our special universal application for working with aligned parallel corpora. Boldface type is applied to those sentences which have no parallel equivalents after automatic alignment.

As a result of estimating at least of three experts, it was determined that about 60% of sentences in our parallel Kazakh-Russian corpus are automatically correctly aligned. The rest of the sentences need to be aligned manually.

In our opinion, such a percentage is connected to the following factors.

1. Not full coincidence by the number of sentences in corpora. In connection with the complexity of syntactic structures of the Kazakh language, some sentences do not correspond on the structure to Russian equivalents and divide to a few sentences.
2. The complexity of dictionary base creation. The basis of the dictionary method lies in the qualitative-designed dictionary. As the Russian and Kazakh languages are in the distant language groups, during the creation of such a dictionary, there are difficulties with accurate translation.
3. The complexity and limitation of using comparative grammar for the Kazakh and Russian languages in our study. The analysis and further development of this approach will allow improving the result of the alignment.
4. The dictionary method does not consider proper nouns. Texts on criminal subject contain a large number of such words, especially in headings.

All these mismatches can significantly affect the results of alignment.

6 Conclusions and further work

Parallel corpora are used for the meeting of different challenges, such as development and setting the machine translation systems, comparative studying of languages, lan-

guage training. Development of such corpora is particularly important for low-resource languages and for pairs of languages related to different groups, for example, for Russian and Kazakh.

The developed parallel Kazakh-Russian corpus is created on the basis of four multilingual news websites of Kazakhstan from which the specialized criminal information was selected. The corpus contains 50410 words from which 25600 relates to Kazakh, and 24800 to Russian.

The corpus is aligned with the use of the specially configurated dictionary and knowledge of POS-tagging of both texts. Additionally, the corpus is provided with the special software application allowing adding specialized information to the corpus. The expert assessment of the automatic alignment is 60% of correctly aligned texts. In the next phase of the study, we plan to classify and analyze the mistakes connected with the alignment of the corpus. For that, we will involve a group of philologists of the Kazakh and Russian languages to the professional analysis and assessment of the results.

The developed aligned Kazakh-Russian parallel corpus can be used as training data for machine translation, identification and extraction of the texts connected with crime and for various NLP tasks.

The following step of our study is greater involvement in a stage of the information alignment using POS-tagging which is limited by the complexity of such full marking for the Kazakh texts now.

7 Acknowledgment

This research is supported by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan (project No. AP05131073 – Methods, models of retrieval and analyses of criminal contained information in semi-structured and unstructured textual arrays).

References

1. Rizun, N., Waloszek, W.: Methodology for Text Classification using Manually Created Corpora-based Sentiment Dictionary. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) – Volume 1: KDIR, pp. 212-220 (2018)
2. Gale, W.A., Church, K.W.: A program for aligning sentences in bilingual corpora. In: ACL'93 29th Annual Meeting, vol. 19(1), pp.75–102. USA (NJ) (1993)
3. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., VíSuchomel, V.: The Sketch Engine: Ten Years On. In: Lexicography, pp. 7-36 (2014)
4. Chapaev, D., Turapbekov, B.: Building Kazakh language open source corpora using wikipedia resources. In: Suleyman Demirel University Bulletin, pp. 153- 160 (2018)
5. Makhambetov, O., Makazhanov, A., Yessenbayev, Z., Matkarimov, B., Sabyrgaliyev, I., Sharafudinov, A.: Assembling the Kazakh Language Corpus. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1022-1033 (2013)

6. Kay, M., Roscheisen, M.: Text translation alignment. Computational Linguistics, 19(1), 121–142 (1993)
7. Fung, P., McKeown, K.: Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. In: Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94), pp. 81–88. Columbia, Maryland, USA (1994)
8. Simard, M., Foster G.F., Isabelle, P., Using cognates to align sentences in bilingual corpora. In: Proceedings of the Fourth International conference on theoretical and methodological issues in Machine translation (TMI 1992), pp. 67–81 (1992)
9. Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., Tron, V.: Parallel corpora for medium density languages. Amsterdam Studies In: The Theory And History Of Linguistic Science Series 4(292), 247 (2007)
10. Sennrich, P., Volk, M.: Iterative, MT-based Sentence Alignment of Parallel Texts. In: Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011), pp. 175-182 (2011)
11. Li, P., Sun, M., Xue, P.:Fast-Champollion: A Fast and Robust Sentence Alignment Algorithm. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 710-718. Beijing, China (2010)
12. Vondricka, P.: Aligning parallel texts with InterText. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). pp. 1875-1879 (2014)
13. Zhumanov, Z., Madiyeva, A., Rakhimova, D.: New Kazakh parallel text corpora with online access. In: Conference on Computational Collective Intelligence Technologies and Applications, pp. 501-508 (2017)
14. Rakhimova, D., Zhumanov, Z.:.Complex Technology of Machine Translation Resources Extension for the Kazakh Language. Advanced Topics in Intelligent Information and Database Systems. Springer International Publishing, Almaty, Kazakhstan (2017)
15. Grabar, N., Kanishcheva, O., Hamon, T.: Multilingual aligned corpus with Ukrainian as the target language. In: SLAVICORP, pp. 53-57 (2018)
16. Harme, J.: Last year but not yesterday? Explaining differences in the locations of Finnish and Russian time adverbials using comparable corpora. In: SLAVICORP, pp. 60-63 (2018)
17. Smith, J. R., Quirk, C., Toutanova, K.: Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In: Proceedings of the Human language Technologies/North American Assosiation for Computational Linguistics, pp. 403-411 (2010)
18. Lewoniewski, W., Węcel, K., Abramowicz, W. Quality and importance of Wikipedia articles in different languages. In International Conference on Information and Software Technologies, pp. 613 – 624 (2016)
19. Rosen, A.: In search of the best method for sentence alignment in parallel texts. In Computer treatment of Slavic and East European languages. Third international seminar, pp. 174-185 (2005)