UBMK'24

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

9th International Conference on Computer Science and Engineering

26-27-28 Ekim (October) 2024 Antalya - Türkiye

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2024)

9th International Conference on Computer Science and Engineering

26-28 Ekim 2024 Akdeniz Üniversitesi Antalya Türkiye
26-28 October 2024 Akdeniz University Antalya Türkiye

IEEE Catalog Number : CFP24L97-CDR

ISBN : 979-8-3503-6587-0

Additional copies may be ordered from:
Curran Associates, Inc
57 Morehouse Lane Red Hook, NY 12571 USA
Phone: (845) 758 0400
Fax: (845) 758 2633
E-mail: curran@proceeding.com

An Online Platform for Uzbek-Russian and Russian-Uzbek Parallel Corpora:

Linguistic Challenges and Prospects Exemplified by A. Kadyri’s Novel “Bygone Days”

Nigmatova Lolakhon Khamidovna
Russian language and literature
Bukhara State University
Bukhara, Uzbekistan
nigmatovalolaxon@gmail.com

Saidova Mokhira Rasulevna
Russian language and literature
Bukhara State University
Bukhara, Uzbekistan
saidova.m549@gmail.com

Djuraeva Zulkhumor Radzhabovna
Russian language and literature
Bukhara State University
Bukhara, Uzbekistan
djuraeva.z.r.17@gmail.com

Sharipov Sokhib Salimovich
Russian language and literature
Bukhara State University
Bukhara, Uzbekistan
Sharipov.S.S.1777@gmail.com

Avezov Sukhrob Sobirovich
Russian language and literature
Bukhara State University
Bukhara, Uzbekistan
1990senigama@gmail.com

***Abstract*— This article explores the theoretical and practical foundations of creating an Uzbek-Russian parallel corpus and developing online platforms for these corpora, based on A.Kadyri’s novel “Bygone days”. The primary objective is to develop an accessible tool for enhancing machine translation and linguistic analysis between Uzbek and Russian at the phrase, sentence, and text levels. The study highlights the distinctive features of the Uzbek-Russian parallel corpora, gathers and prepares parallel texts from the novel, and focuses on the platform’s functionalities like search, alignment, comparison, and translation. The platform’s efficiency is assessed through experiments and user experience, and linguistic analyses such as idiom translation and linguistic realities are conducted. The research employs various methodologies including classification, descriptive, comparative, distributional, translation methods, and corpus text analysis. The novelty lies in the creation of the Uzbek-Russian parallel corpus and the development of a platform for its utilization, emphasizing its unique contribution to language learning and translation studies. This innovative platform significantly enhances research efficiency, advancing the translation of idiomatic expressions and linguistic realities between Uzbek and Russian.**

***Keywords*— uzbek-russian parallel corpus, machine translation, linguistic analysis, online platform, idiomatic expressions, linguistic realities.**

I. Introduction

In the era of digitalization, the development of multilingual computational tools has become essential for advancing linguistic research and improving translation accuracy. Parallel corpora serve as crucial resources in this endeavor, enabling detailed analysis and comparison of texts across languages. By facilitating the alignment and translation of complex linguistic structures, these tools contribute significantly to the fields of language learning, translation studies, and cultural exchange. This paper presents an innovative approach to creating and utilizing parallel corpora for Uzbek and Russian, highlighting the potential for enhancing cross-linguistic understanding and communication.

II. Literature review

The need for organizing and systematizing accumulated information predates computers, with early efforts including cards, glossaries, and encyclopedias. Until the mid-20th century, computer technologies were mainly used in exact sciences [5 pp; 54]. However, the intersection of disciplines led to the emergence of computational and corpus linguistics, fully established by the 1990s [11 pp; 5]. Despite its historical roots in concordances from the 13th century [16 pp; 165], corpus linguistics is now seen as a modern field, leveraging computer technologies for linguistic research and maximizing the benefits of new information processing methods.

The study of theoretical and methodological problems in linguistic support began in the 1960s, recognizing the importance of information retrieval languages in transitioning computers to the “logical-linguistic” stage. Key contributions to natural language processing were made by scholars such as S.Daniel, Zhuravsky, James Martin (Speech and language processing), and M.A.Mohri (Finite-state transducers in language and speech processing). The first parallel corpus, the Brown Corpus, was developed by W.Francis and H.Kučera in 1960. Significant works on corpus technologies in language teaching include D.Sinclair’s studies. Research on Russian parallel corpora by D.O.Dobrovolsky, Sichinava, Yu.Tao, V.Zakharov, S.A. Manik, and A.A. Kokoreva is notable. In domestic linguistics, contributions by A.G. Gilemshina, A.K. Khudoyberdiev, N.Z. Abdurakhmanova, Sh.M. Khamroeva, M.A. Abzhalova, N.A. Ataboeva and others have advanced the field. The text and its components are crucial in creating parallel corpora, with Uzbek scholars like A.Mamajanov, N.M.Turniyozov, E.Qilichev, M.Hakimov, and others significantly contributing to text linguistics, which is vital for data collection and enrichment in corpora.

Recent research highlights advancements in parallel corpora and machine translation. A.Fan discuss non-English-centric multilingual machine translation. A.Sirchina explores corpus linguistics methods in translation. D.Dobrovolsky examines parallel texts in lexical semantics. M.R.Costa-Jussa and J.Fonollosa propose a character-based neural machine translation approach. J.Hu introduce XTREME, a multilingual benchmark for cross-lingual generalization. Firat, Cho, and Bengio present a multilingual neural translation approach with shared attention.

Other notable works include M.Artetxe on unsupervised neural translation, Ruder, Vulić, and Søgaard on cross-lingual word embedding models, Yen, Huang, and Chen on

English-Chinese bilingual word representations, and H.Schwenk on WikiMatrix, a large multilingual parallel corpus. S.Musurmankulova discusses the formation of the Uzbek-Turkish parallel corpus. B.Allaberdiev present a dataset for Uzbek-Kazakh machine translation.

These studies underscore the importance of parallel corpora in enhancing translation quality and advancing linguistic research.

III. Data Collection and Alignment Process for the Uzbek-Russian Parallel Corpus

The creation and application of parallel corpus platforms is a priority in modern linguistics, enabling in-depth analysis and comparison of texts across languages. These platforms reveal linguistic features reflecting cultural, historical, and sociolinguistic aspects. Essential for translating literary, scientific, and official documents between Uzbek and Russian, these tools address the long-standing linguistic ties between Uzbekistan and Russia. Parallel corpora, collections of aligned texts in multiple languages [7 pp; 102] [10 pp; 15], are crucial in machine translation and linguistic studies, aiding in the development of accurate translation models [1 pp; 8] [9 pp; 4415]. The concept emerged mid-20th century, gaining momentum with computational linguistics in the 80s-90s [19 pp; 42]. Early projects, like the Europarl corpus, demonstrated their value. Today, numerous parallel corpora exist, continuously expanding through digital resources, supporting advancements in automatic translation.

Key scientific development vectors in this field include:

1. Expanding language coverage: while large corpora exist for popular languages, many rare languages remain underrepresented, necessitating the creation of corpora for these languages [20 pp; 32] [14 pp; 3];

2. improving alignment methods: correct alignment of source text sentences with their translations is crucial for training machine translation models, highlighting the need for new methods and algorithms [4 pp; 722] [17 pp; 581];

3. integration with other linguistic resources: combining parallel corpora with lexicographic, grammatical, and other resources can enhance research capabilities [12 pp; 99] [13 pp; 4].

For the development of Uzbek-Russian parallel corpora, researchers must address linguistic tasks, such as accounting for the structural, grammatical, lexical, and stylistic features of each language, as well as the cultural and sociolinguistic aspects in translation. This approach not only aids language study but also reveals new research opportunities.

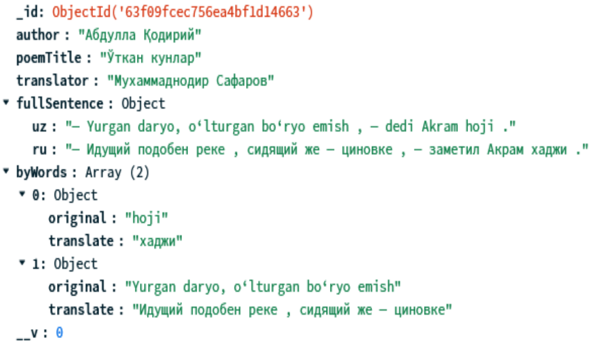
Our system, developed using Visual Studio Code, JavaScript, ReactJS, NodeJS, and ExpressJS, operates on both local and network levels. The web application, based on ReactJS, uses MongoDB for database management and storage. This platform can be used for language learning, translation, and philological studies, serving as a database and toolset for educational purposes.

The primary data for this study were sourced from the original Uzbek text of A.Kadyri's “Bygone days” and its corresponding Russian translation by Muhammadnadir Safarov, published in 2009. The corpus consists of 247,600 words and 21,346 sentences. The data collection was executed manually, a necessity stemming from the limitations of existing automatic alignment platforms which do not support the Uzbek language.

Current automatic alignment platforms, such as ParaConc, Bitext2tmx, Moses, and Sketch Engine, lack support for Uzbek, necessitating a manual alignment process. This lack of support presents a significant challenge in aligning texts from less commonly supported languages, thus requiring a labor-intensive manual approach.

The alignment process was conducted in two distinct stages to ensure high precision. The initial phase involved aligning the texts at the sentence level. Each sentence from the original Uzbek text was manually matched with its corresponding sentence in the Russian translation. This stage aimed to create a broad alignment framework, setting the foundation for more granular alignment. Following the primary alignment, a more detailed alignment was performed at the phrase and word levels. This secondary alignment refined the initial sentence-level alignment, ensuring that phrases and individual words within each sentence were accurately matched. This process can be described as an alignment within an alignment, where finer details are meticulously aligned within the broader sentence-level framework.

Upon completion of the manual alignment process, the aligned data were manually uploaded to MongoDB Atlas, a scalable and flexible database management system.



```
  _id: ObjectId('63f09fcec756ea4bf1d14663')
  author : "Абдулла Қодирий"
  poemTitle : "Ўткан кунлар"
  translator : "Мухаммаднодир Сафаров"
▾ fullSentence : Object
    uz : "– Yurgan daryo, o‘lturgan bo‘ryo emish , – dedi Akram hoji ."
    ru : "– Идущий подобен реке , сидящий же – циновке , – заметил Акрам хаджи ."
▾ byWords : Array (2)
  ▾ 0: Object
      original : "hoji"
      translate : "хаджи"
  ▾ 1: Object
      original : "Yurgan daryo, o‘lturgan bo‘ryo emish"
      translate : "Идущий подобен реке , сидящий же – циновке"
  __v : 0
```

Fig.1. Data structure

The developed database consists of 8 interconnected entities:

- «user»: _id, firstName, lastName, userPhoneNumber, regSelectRole, userEmail, userPassword, userOrganization, userPositionSelect;
- «corpus»: _id, authorID, poemID, translatorID, fullSentenceID, byWordsID;
- «author»: _id, name_full, poemID, translatorID;
- «translator»: _id, name_full, authorID, poemID, translated_poemID;
- «fullSentence»: _id, langUZ, langRU, authorID, translatorID, poemID, translated_poemID;
- «poem»: _id, authorID, translatorID, fullSentenceID, poemTitle;
- «byWords»: _id, fullSentenceID, poemID;
- «translated_poem»: _id, translatorID, authorID, poemID, poemTitle.

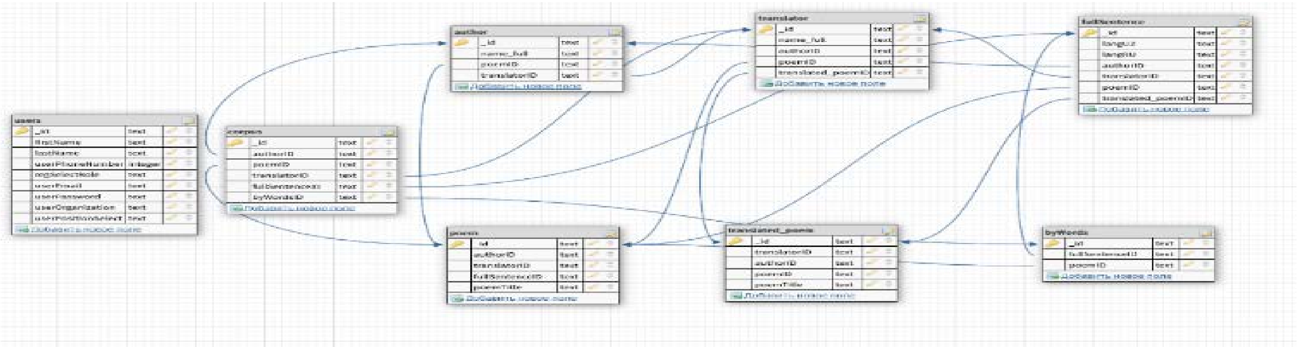


Fig 2. Database structure

IV. BACKEND IMPLEMENTATION

The backend implementation of the developed platform was achieved through the creation of APIs utilizing NodeJS and ExpressJS to handle requests directed towards the MongoDB database, ensuring efficient data management and retrieval processes.

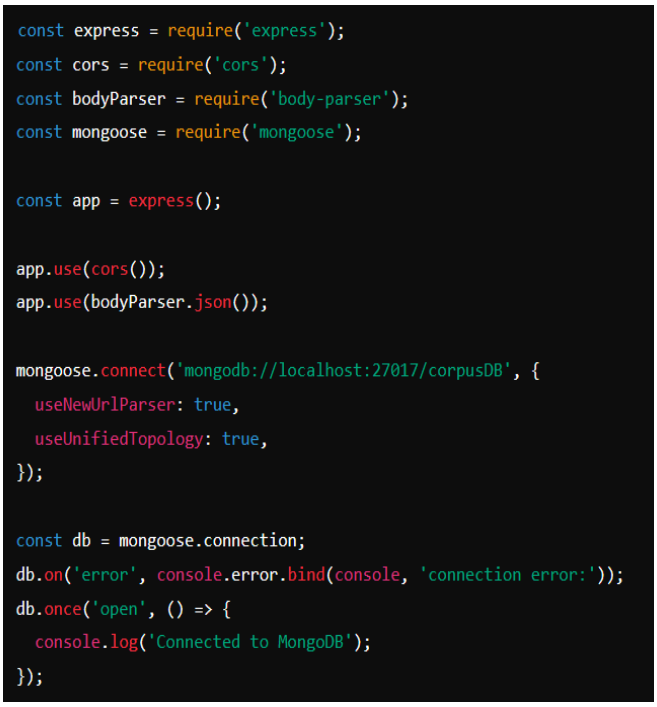
Server-side logic:

NodeJS: provides a scalable, high-performance environment for managing user interactions and database queries. Its non-blocking, event-driven architecture handles multiple simultaneous requests efficiently [6 pp; 67].

API structure:

ExpressJS: simplifies the creation of server-side code, handling routing, middleware integration, and HTTP requests and responses [6 pp; 68].

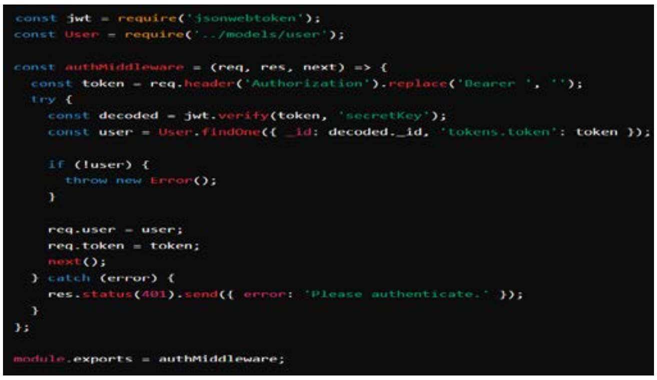
CORS middleware: enables controlled access to resources from different origins, enhancing security by allowing only specified domains to interact with the server.



```
const express = require('express');
const cors = require('cors');
const bodyParser = require('body-parser');
const mongoose = require('mongoose');

const app = express();

app.use(cors());
app.use(bodyParser.json());

mongoose.connect('mongodb://localhost:27017/corpusDB', {
  useNewUrlParser: true,
  useUnifiedTopology: true,
});

const db = mongoose.connection;
db.on('error', console.error.bind(console, 'connection error:'));
db.once('open', () => {
  console.log('Connected to MongoDB');
});
```

Fig.3. Middleware integration

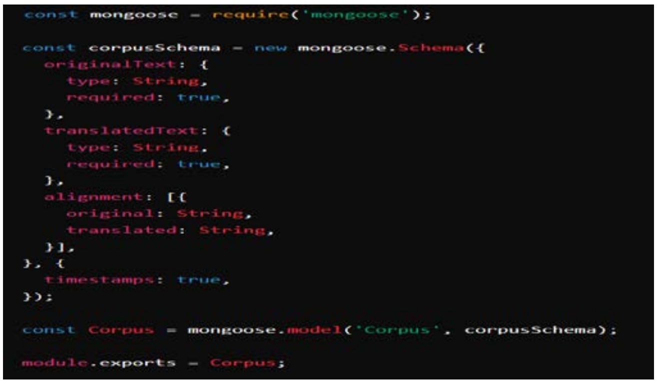
Validates user credentials, generates secure tokens, and manages user sessions to ensure only authorized access to data.



```
const jwt = require('jsonwebtoken');
const User = require('../models/user');

const authMiddleware = (req, res, next) => {
  const token = req.header('Authorization').replace('Bearer ', '');
  try {
    const decoded = jwt.verify(token, 'secretKey');
    const user = User.findOne({ _id: decoded._id, 'tokens.token': token });

    if (!user) {
      throw new Error();
    }

    req.user = user;
    req.token = token;
    next();
  } catch (error) {
    res.status(401).send({ error: 'Please authenticate.' });
  }
};

module.exports = authMiddleware;
```

Fig.4. Authentication middleware

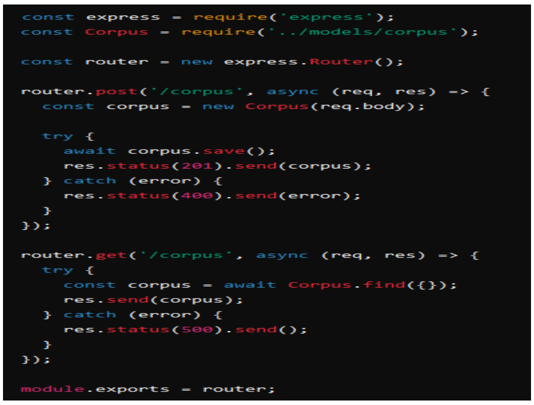
MongoDB: utilizes a flexible schema to store and manage large volumes of text data efficiently. Collections are optimized for rapid access and retrieval.



```
const mongoose = require('mongoose');

const corpusSchema = new mongoose.Schema({
  originalText: {
    type: String,
    required: true,
  },
  translatedText: {
    type: String,
    required: true,
  },
  alignment: [{
    original: String,
    translated: String,
  }],
}, {
  timestamps: true,
});

const Corpus = mongoose.model('Corpus', corpusSchema);

module.exports = Corpus;
```

Fig.5. Database design

Encapsulate core logic for handling specific types of requests, such as CRUD operations for corpus data.



```
const express = require('express');
const Corpus = require('../models/corpus');

const router = new express.Router();

router.post('/corpus', async (req, res) => {
  const corpus = new Corpus(req.body);

  try {
    await corpus.save();
    res.status(201).send(corpus);
  } catch (error) {
    res.status(400).send(error);
  }
});

router.get('/corpus', async (req, res) => {
  try {
    const corpus = await Corpus.find({});
    res.send(corpus);
  } catch (error) {
    res.status(500).send();
  }
});

module.exports = router;
```

Fig 6. Controllers

The backend, with NodeJS, ExpressJS, and MongoDB, provides a robust foundation for the platform, facilitating advanced linguistic research and translation studies by ensuring efficient data management and retrieval.

V. Frontend Implementation

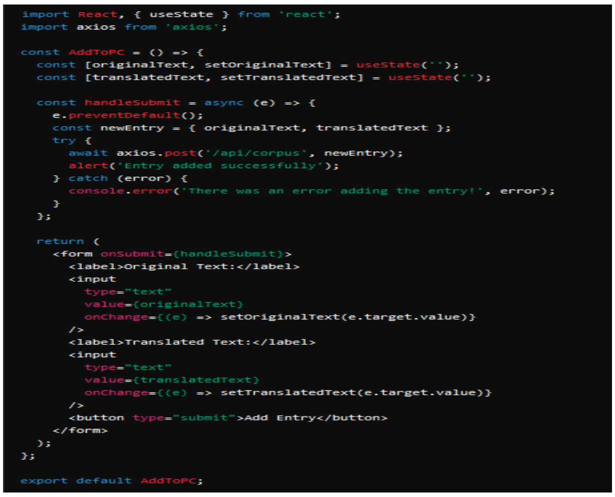
The frontend implementation of the developed platform utilizes ReactJS to create a dynamic and responsive user interface. The frontend components handle user interactions, display data, and communicate with the backend APIs to ensure a seamless user experience.

ReactJS components:

AddToPC component:

- manages the addition of new parallel corpus entries.

- handles user input, form submission, and communicates with the backend to store the data in MongoDB.



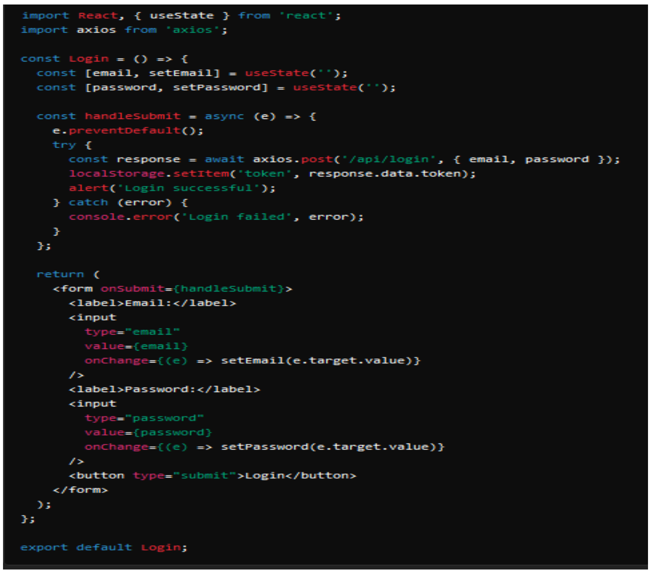
```
import React, { useState } from 'react';
import axios from 'axios';

const AddToPC = () => {
  const [originalText, setOriginalText] = useState('');
  const [translatedText, setTranslatedText] = useState('');

  const handleSubmit = async (e) => {
    e.preventDefault();
    const newEntry = { originalText, translatedText };
    try {
      await axios.post('/api/corpus', newEntry);
      alert('Entry added successfully');
    } catch (error) {
      console.error('There was an error adding the entry!', error);
    }
  };

  return (
    <form onSubmit={handleSubmit}>
      <label>Original Text:</label>
      <input
        type="text"
        value={originalText}
        onChange={(e) => setOriginalText(e.target.value)}
      />
      <label>Translated Text:</label>
      <input
        type="text"
        value={translatedText}
        onChange={(e) => setTranslatedText(e.target.value)}
      />
      <button type="submit">Add Entry</button>
    </form>
  );
};

export default AddToPC;
```

Fig 7. AddToPC component

Login component:

Manages user authentication;

Handles user input for email and password, form submission, and communicates with the backend to validate user credentials.



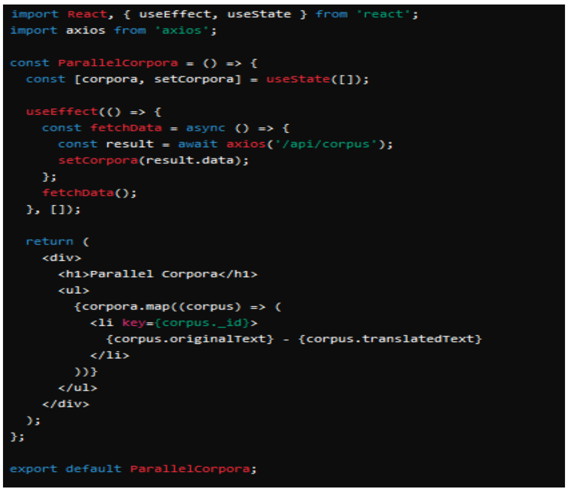
```
import React, { useState } from 'react';
import axios from 'axios';

const Login = () => {
  const [email, setEmail] = useState('');
  const [password, setPassword] = useState('');

  const handleSubmit = async (e) => {
    e.preventDefault();
    try {
      const response = await axios.post('/api/login', { email, password });
      localStorage.setItem('token', response.data.token);
      alert('Login successful');
    } catch (error) {
      console.error('Login failed', error);
    }
  };

  return (
    <form onSubmit={handleSubmit}>
      <label>Email:</label>
      <input
        type="email"
        value={email}
        onChange={(e) => setEmail(e.target.value)}
      />
      <label>Password:</label>
      <input
        type="password"
        value={password}
        onChange={(e) => setPassword(e.target.value)}
      />
      <button type="submit">Login</button>
    </form>
  );
};

export default Login;
```

Fig 8.. Login component

ParallelCorpora component:

Displays the list of parallel corpora entries;

Retrieves data from the backend and renders it in a user-friendly format.



```
import React, { useEffect, useState } from 'react';
import axios from 'axios';

const ParallelCorpora = () => {
  const [corpora, setCorpora] = useState([]);

  useEffect(() => {
    const fetchData = async () => {
      const result = await axios('/api/corpus');
      setCorpora(result.data);
    };
    fetchData();
  }, []);

  return (
    <div>
      <h1>Parallel Corpora</h1>
      <ul>
        {corpora.map((corpus) => (
          <li key={corpus._id}>
            {corpus.originalText} - {corpus.translatedText}
          </li>
        ))}
      </ul>
    </div>
  );
};

export default ParallelCorpora;
```

Fig.9. ParallelCorpora component

VI. Platform Description

The platform we have developed comprises two primary components: the “Parallel corpus analysis page” and the “Parallel corpus upload window”. Each component is designed to facilitate detailed linguistic analysis and efficient data management within the parallel corpus database.

Parallel corpus analysis page: the parallel corpus analysis page consists of a search bar and several filters, including corpus language selection, author selection, work or document selection, and translator selection. These filters enable users to narrow down their searches to specific segments of the corpus, thereby allowing precise and targeted retrieval of information. If the filters are not utilized, the search operates across the entire corpus, ensuring that all relevant data is considered. This page is integral for conducting detailed linguistic inquiries, as it helps users to efficiently locate and analyze text pairs within the corpus. By providing multiple filtering options, the platform supports a wide range of search criteria, making it versatile for various types of linguistic research.

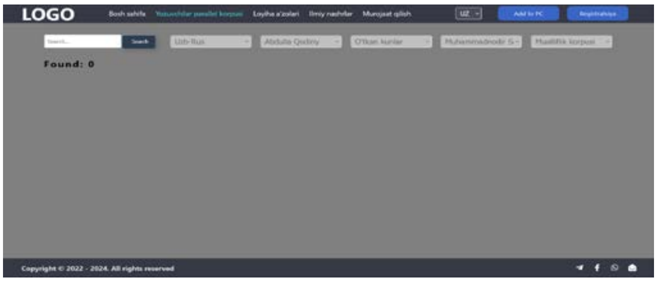


Fig.10. Parallel corpus analysis page

Parallel corpus upload window: the Parallel corpus upload window is designed to allow users to upload text data into the parallel corpus database. This component supports manual alignment and semantic tagging of the uploaded texts. Users can input the original and translated texts and then perform manual alignment to ensure that each segment of text is properly matched between the two languages. This two-step alignment process involves an initial broad alignment at the sentence level, followed by a more detailed alignment at the phrase or word level. Semantic tagging is also manually applied to enhance the corpus's usability for linguistic analysis. This manual process, though labor-intensive, ensures a high level of accuracy in the alignment and tagging of texts, which is crucial for the integrity of linguistic research. The Parallel Corpus Upload Window also includes features for metadata entry, allowing users to add relevant information about the texts, such as the author, publication date, and source. This metadata is essential for organizing and contextualizing the data within the corpus.

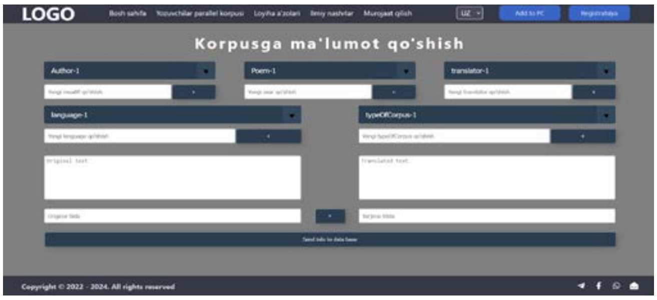


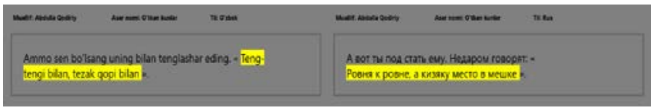
Fig.11. Parallel corpus upload window

Web platform, with their user-friendly data processing tools, allow even inexperienced users to manage and manipulate server databases efficiently, performing tasks such as adding, deleting, searching, and processing data.

Existing online resources for parallel corpus creation include platforms like ParaConc (paid), Bitext2tmx (limited functionality), Moses (limited functionality), and Sketch Engine (lacks Uzbek language support) [2 pp; 94] [15 pp; 4]. However, these systems have several drawbacks: closed formats, rigid output formats, limited functionality, paid versions, lack of necessary languages, and no tagging capabilities [8 pp; 6]. Therefore, developing a user-friendly and versatile information resource for creating Uzbek-Russian parallel corpora remains a pertinent task.

The novel "Bygone days" by Abdulla Kadyri is a cornerstone of Uzbek literature, captivating readers with its gripping plot, unique language, and expressive techniques. It addresses timeless themes of love, friendship, honor, patriotism, family relationships, and generational conflicts. Although Kadyri does not explicitly mention the Jadid movement, the protagonist, Otabek, embodies the progressive youth committed to the bright future of his people. His dedication to his ideals, readiness for sacrifices, and love for the heroine, Kumush, empower him further. Kadyri skillfully portrays the lifestyle of the people and the historical events that significantly impacted their lives during those bygone days.

Translating idiomatic expressions is challenging due to their figurative nature and cultural richness. Accurate translation requires deep knowledge of both cultures and languages to convey the correct meaning and stylistic nuance [18 pp; 17]. Despite the difficulties, various methods and techniques enable translators to creatively and adequately render idioms, ensuring high-quality and comprehensible translations. This work aims to explore and analyze methods for translating idiomatic expressions.



Muallif: Abdulla Qodiriy Asar nomi: O'tkan kunlar Til: O'zbek

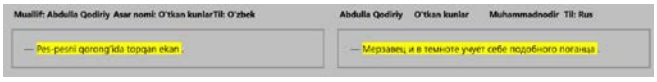
Ammo sen bo'lsang uning bilan tenglashar eding. « Teng-tengi bilan, tezak qopi bilan ».

Muallif: Abdulla Qodiriy Asar nomi: O'tkan kunlar Til: Rus

А вот ты под стать ему. Недаром говорят: « Ровня к ровне, а кизяку место в мешке ».

Fig.12. Example of translation of idiomatic expressions

The half-joking words of the maid, Toybeka, about a suitable groom for Kumush, a "young guest", irritate the latter, who mockingly suggests that the maid herself marry him. The maid responds with the proverb: "Teng-teng bilan, tezak qopi bilan" (Like seeks like, and dung belongs in a sack) [3 pp; 32]. M.Safarov translates this literally: "Ровня к ровне, а кизяку место в мешке" [21 pp; 38]. At first glance, this seems like a successful translation: the meaning is accurately conveyed, and it sounds proverb-worthy. Given the proverb's strong connection to Uzbek daily life, the translation seems appropriate. However, in a footnote, the word "кизяк" (dung) is explained as "навоз" (manure), which undermines the translation's success. The term "кизяк" refers to manure dried into bricks used as fuel, not just manure itself. These dried bricks are often stored in sacks. Due to the incorrect explanation, Russian-speaking readers might be puzzled, unable to reconcile "маниre" with "sack", since storing manure in a sack is unusual.



Muallif: Abdulla Qodiriy Asar nomi: O'tkan kunlarTil: O'zbek

— Pes-pesni qorong'ida topqan ekan .

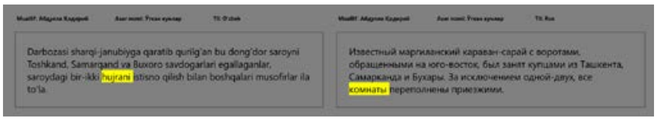
Abdulla Qodiriy O'tkan kunlar Muhammadnodir Til: Rus

— Мерзавец и в темноте учует себе подобного поганца .

Fig.13. Example of translation of idiomatic expressions

The original Uzbek idiom "Pes-pesni qorong'ida topqan ekan" (A person with vitiligo will find another with vitiligo even in the dark) means that similar people find each other even in difficult circumstances. The Russian translation "Мерзавец и в темноте учует себе подобного поганца" (A scoundrel will sniff out a fellow scoundrel even in the dark) does not fully convey the original meaning, although it retains the idea of similar individuals finding each other. The translator used partial equivalence, preserving some core elements of the original idiom's meaning. This transformation makes the idiom understandable for the target audience but slightly distorts the original sense.

Translating linguistic realities between Uzbek and Russian is challenging due to different cultural, historical, and sociolinguistic contexts. Uzbek terms like "hujra" (cell), "ko'rpa" (quilt), and "varaqi" (leaflet) reflect cultural specifics without direct Russian equivalents. Descriptive methods or calques are often needed. This study analyzes these translation methods using Abdulla Kadyri's novel "Bygone days" and its Russian translation by M.Safarov, employing a parallel corpus of both languages.



Muallif: Абдулла Кадырий Asar nomi: Ўтган кунлар Til: O'zbek

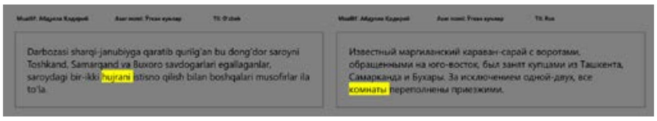
Darbozasi sharqi-janubiyga qaratib qurilg'an bu dong'dor saroyni Toshkand, Samarqand va Buxoro savdogarlari egallaganlar, saroydagi bir-ikki hujrani istisno qilish bilan boshqalari musofirlar ila to'la.

Muallif: Абдулла Кадырий Asar nomi: Ўтган кунлар Til: Rus

Известный маргиланский караван-сарай с воротами, обращенными на юго-восток, был занят купцами из Ташкента, Самарканда и Бухары. За исключением одной-двух, все комнаты переполнены приезжими.

Fig.14. Example of translation of linguistic realities

The word "hujra" originates from Arabic and has two meanings. The first meaning is "a small room", while the second meaning, "a cell", has historical connotations, describing a living space in a madrasa, primarily used by teachers and students in Central and Western Asia. In the work, "hujra" refers to a small room in a caravanserai where Atabek stayed during his travels in Margilan. Today, this word has fallen out of active use in the Uzbek language and retains only its second meaning – “a small room in a madrasa, caravanserai, or mosque”. The author’s use of this word helps to recreate the historical setting.



Муаллиф: Абдулла Қодирий Асар номи: Ўтган кунлар Til: O'zbek

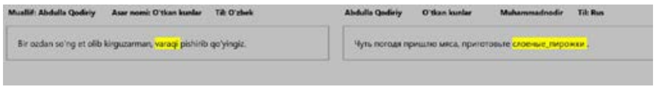
Darbozasi sharqi-janubiyga qaratib qurilg'an bu dong'dor saroyni Toshkand, Samarqand va Buxoro savdogarlari egallaganlar, saroydagi bir-ikki hujrani istisno qilish bilan boshqalari musofirlar ila to'la.

Муаллиф: Абдулла Қодирий Асар номи: Ўтган кунлар Til: Rus

Известный маргиланский караван-сарай с воротами, обращенными на юго-восток, был занят купцами из Ташкента, Самарканда и Бухары. За исключением одной-двух, все комнаты переполнены приезжими.

Fig.15. Example of translation of linguistic realities

The next word, “ko‘rpa”, was mistakenly replaced with “ko‘rpacha”. This led to a distortion of meaning, although the cultural and historical features remained the same. This occurred due to the use of transliteration as a translation method. “Ko‘rpa” in Russian corresponds to “одеяло, покрывало” (blanket, cover), while “ko‘rpacha” has no direct equivalent but semantically differs from “ko‘rpa”, meaning “a narrow silk quilt filled with cotton, intended for sitting”.



Muallif: Abdulla Qodiriy Asar nomi: O'tkan kunlar Til: O'zbek

Bir ozdan so'ng et olib kirguzarman, varaqi pishirib qo'yingiz.

Abdulla Qodiriy O'tkan kunlar Muhammadnodir Til: Rus

Чуть погодя пришлю мяса, приготовьте слоеные_пирожки .

Fig.16. Example of translation of linguistic realities

The word “varaqi” is Uzbek and refers to a type of samsa with various shapes and fillings, usually meat, commonly prepared in the Ferghana Valley for special occasions. It derives from the word “varaq” (leaf), describing the layered texture of the pastry. In Russian, it is translated as “слоеные пирожки” (layered pies), which does not fully convey the cultural specifics of “varaqi.” This example illustrates the challenges of translating culturally specific words, as seen in A. Kadyri’s novel “Bygone days.” The Uzbek-Russian parallel corpus aids in analyzing and enhancing the understanding of such nuances in translation.

VII. Conclusion

The development and implementation of an online platform for Uzbek-Russian and Russian-Uzbek parallel corpora, exemplified by A. Kadyri’s novel “Bygone days”, significantly advance computational linguistics. This study addresses the challenges of translating idiomatic expressions and cultural terms, emphasizing detailed linguistic analysis and precise alignment methods. The platform utilizes modern technologies such as Visual Studio Code, JavaScript, ReactJS, NodeJS, and MongoDB, creating a user-friendly system for managing linguistic data. It highlights the structural, grammatical, lexical, and stylistic characteristics of both languages, essential for accurate translation. The platform serves as a valuable educational and research tool, enhancing language learning and translation studies. By facilitating the study of cross-linguistic nuances, it contributes to a deeper understanding of cultural aspects in Uzbek and Russian. The research also opens new avenues for expanding language coverage, improving alignment methods, and integrating parallel corpora with other linguistic resources, thus advancing translation models and linguistic research. This innovative platform marks a significant milestone, offering robust tools for machine translation and linguistic analysis, and paving the way for future developments in computational linguistics and cross-linguistic understanding.

References

[1] A.Fan et al. Beyond english-centric multilingual machine translation //Journal of Machine Learning Research. – 2021. – T. 22. – №. 107. – Pp. 1-48.

[2] A.Khusainov et al. First Results of the TurkLang-7 Project: Creating Russian-Turkic Parallel Corpora and MT Systems //CMCL. – 2020. – Pp. 90-101.

[3] A.Qodiriy. O ‘tkan kunlar // Universitet nashriyoti: Toshkent. – 2018.

[4] A.Sirchina. Possibilities of using corpus linguistics methods in translation. Economics and society. – No. 4-2 (23). – 2016. – pp. 720-724.

[5] A.Z.Yen, H.H.Huang, H.H.Chen. Learning English–Chinese bilingual word representations from sentence-aligned parallel corpus //Computer Speech & Language. – 2019. – T. 56. – Pp. 52-72.

[6] B.Basumatary, N.Agnihotri. Benefits and Challenges of Using NodeJS //International Journal of Innovative Research in Computer Science & Technology. – 2022. – T. 10. – №. 3. – C. 67-70.

[7] D.Dobrovolsky. Corpus of parallel texts in the study of lexical semantics. In computational linguistics and intellectual technologies: proceedings of the international conference. “Dialogue”. – 2004. – Pp. 98-122.

[8] H.Schwenk et al. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia //arXiv preprint arXiv:1907.05791. – 2019.

[9] J.Hu et al. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation //International Conference on Machine Learning. – PMLR, 2020. –PpC. 4411-4421.

[10] L.Gilardi L, C.F.Baker. Learning to align across languages: Toward multilingual framenet //Proceedings of the international FrameNet workshop. – 2018. – C. 13-22.

[11] M.Artetxe. et al. Unsupervised neural machine translation //arXiv preprint arXiv:1710.11041. – 2017.

[12] M.Kopotev. On some consequences of corpus linguistics for general language theory. Philological class. – Vol. 26. – No. 2. – 2021. – pp. 90-102.

[13] M.R.Costa-Jussa, J.A.Fonollosa. Character-based neural machine translation //arXiv preprint arXiv:1603.00810. – 2016.

[14] O.Fira, K.Cho, Y.Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism //arXiv preprint arXiv:1601.01073. – 2016.

[15] R.Yeshpanov, A.Polonskaya, H.A.Varol. KazParC: Kazakh Parallel Corpus for Machine Translation //arXiv preprint arXiv:2403.19399. – 2024.

[16] S.Musurmankulova. Theoretical Basis of the Formation of Uzbek-Turkish Parallel Corpus //Central Asian Journal of Literature, Philosophy and Culture. – 2023. – T. 4. – №. 12. – Pp. 163-170.

[17] S.Ruder, I.Vulić, A.Søgaard. A survey of cross-lingual word embedding models //Journal of Artificial Intelligence Research. – 2019. – T. 65. – Pp. 569-631.

[18] S.Sheremetyeva. Controlled Authoring In A Hybrid Russian-English Machine Translation System //Proceedings of the 3rd Workshop on Hybrid Approaches to Machine Translation (HyTra). – 2014. – Pp. 15-20.

[19] V.Zakharov et al. Corpus methods and semantic fields: the concept of empire in english, russian and czech //Proceedings of the III International Conference on Language Engineering and Applied Linguistics, Saint Petersburg, Russia. – 2020.

[20] V.Zakharov, S.Bogdanova. Corpus linguistics: a textbook for humanities students. Irkutsk State Linguistic University. – 2011. – 181 pages.

[21] А.Кадыри. Минувшие дни // Государственное издательство художественной литературы РУз. – 2009. – 371 с.