IEEE
TÜRKİYE SECTION

UBMK'25

Bildiriler Kitabı
Proceedings

Editör Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye abv

IEEE

IEEE TÜRKİYE SECTION

İSTANBUL TEKNİK ÜNİVERSİTESİ 1773

GAZİ ÜNİVERSİTESİ 1926 95. yıl

1 9 9 6 ATILIM ÜNİVERSİTESİ

UBMK'25

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

10th International Conference on Computer Science and Engineering

17-18-19 Eylül (September) 2025 İstanbul - Türkiye

Media type	Part Number	ISBN	Online ISSN
XPLORE COMPLIANT	CFP25L97-ART	979-8-3315-9975-1	2521-1641
CD-ROM	CFP25L97-CDR	979-8-3315-9974-4	

10. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2025)

10th International Conference on Computer Science and Engineering

17-18-19 Eylül 2025 -İstanbul-Türkiye
17-18-19 September 2025 - İstanbul-Türkiye

IEEE Catalog Number : CFP25L97-ART

ISBN : 978-8-3315-9975-1

Additional copies mey 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

UBMK’2025’ye Hoşgeldiniz
Welcome to UBMK’2025

Sevgili Katılımcılar:

UBMK uluslararası nitelikli konferans serisi, 1990 yılından beri düzenli olarak yapılmakta olan Bilgisayar Mühendisliği Bölüm Başkanları toplantılarında alınan bir kararla on yıl önce başlamıştır. Konferansın 10.su IEEE-UBMK-2025 bu yıl 17-18-19 Eylül, 2025 günlerinde İstanbul Teknik Üniversitesinin ev sahipliğinde düzenlemiştir.

IEEE-UBMK-2025 konferansına bu yıl Almanya, Amerika Birleşik Devletleri, Azerbaycan, Fransa, Irak, İngiltere, İsveç, İtalya, Kanada, Kazakistan, Kırım, Kırgızistan, Rusya, Özbekistan, Tataristan, Taylant, Ürdün ve Türkiye’den 610 dolayında bildiri gönderilmiş ve bu bildiriler Türk ve yabancı 250 hakem tarafından değerlendirilmiştir.

Her bildiri en az iki hakem tarafından incelenmiş ve uzlaşma olmadığı durumlarda üçüncü bir hakemin değerlendirmesine başvurulmuştur. Bildiri başına düşen ortalama hakemlik 2,3 olmuştur. Bu değerlendirmelerin sonunda 327 bildirinin sözlü olarak sunulması uygun bulunmuştur. Kabul edilen ve sunulan bildiriler içerik ve kalite ölçünlerini sağlaması durumunda IEEE Xplore’da yayımlanacaktır.

Konferans çalışmalarında, Bilgisayar Mühendisliği Bölüm Başkanları Danışma Kurulu olarak görev almışlardır. Bildirilerin değerlendirilmesi Bilim Kurulu üyeleri tarafından yapılmıştır. Konferansın düzenlenmesi ise Yürütme Kurulunun önerileri doğrultusunda, Düzenleme Kurulu tarafından yapılmıştır.

Son olarak, konferansın başarılı bir şekilde yürütülmesi için tüm olanaklarını sunan İstanbul Teknik Üniversitesi Rektörü Sayın Prof. Dr. Hasan Mandal’a teşekkür ediyoruz. Ayrıca Düzenleme Kuruluna, bildirileri titizlikle değerlendiren Bilim Kurulu Üyelerine ve değerli araştırmalarının sonuçlarını bilişim camiası ile paylaşan bildiri sahiplerine teşekkürlerimizi iletiriz.

Prof. Dr. Eşref ADALI
UBMK-2025 Konferans Başkanı ve Bildiri Kitabı Editörü

Dear Participants:

The UBMK international conference series started nine years ago with a decision taken at the Computer Engineering Department Heads (BMBB) meetings, which have been held regularly since 1990. The 10th edition of the conference, UBMK'25, was held this year on October 17-18-19, 2025, hosted by İstanbul Technichal University.

This year, approximately 610 papers were submitted to the IEEE-UBMK-2025 conference from Germany, the United States, Azerbaijan, France, Iraq, the United Kingdom, Sweden, Italy, Canada, Kazakhstan, Crimea, Kyrgyzstan, Russia, Uzbekistan, Tatarstan, Thailand, Jordan, and Turkey, and these papers were evaluated by 250 Turkish and foreign referees.

Each paper was evaluated at least by two referees, and in cases where there was no consensus, a third referee was consulted. At the end of these evaluations, 327 papers were accepted for oral presentation. Accepted and presented papers will be submitted for inclusion into IEEE Xplore subject to meeting IEEE Xplore's scope and quality requirements.

During the conference, Heads of Information Engineering Departments took part in the Advisory Board. The evaluation of the papers was made by the members of the Scientific Committee. The conference was organized by the Organizing Committee in line with the recommendations of the Executive Committee.

Finally, we would like to thank İstanbul Technical University Rector Prof. Dr. Hasan Mandal for his continued support for the success of the conference. In addition, we would like to thank the Organizing Committee, the Scientific Committee Members who carefully evaluated the papers, and the owners of the papers who shared the results of their valuable research with the informatics community.

Prof. Dr. Esref ADALI
UBMK'25 Conference Chair and Proceedings Editor

Düzenleyenler / Organizer

Destekleyenler / Sponsors

VBT

YapıKredi
Teknoloji

NEOVA
SİGORTA

Orion
Innovation

arabam com

idea
TEKNOLOJİ ÇÖZÜMLERİ

PROTEL

fuzul VENTURES

softtech

KURUKAHVECİ
KURUKAHVECİ
MEHMET EFENDİ

ÜSKÜDAR BELEDİYE BAŞKANLIĞI
ÜSKÜDAR
BELEDİYESİ

Stages of Creating an Uzbek-English Parallel Corpus and Principles of Selecting a Linguistic Base

Botir Elov Boltayevich
Tashkent State University of Uzbek language and literature
Tashkent, Uzbekistan
dotsentelov@navoiy-uni.uz

Ruhillo Alaev Habibovich
National University of Uzbekistan named after Mirzo Ulugbek
Tashkent, Uzbekistan
alayev_r@nuu.uz

Marufjon Amirkulov Alikulovich
Tashkent State University of Uzbek language and literature
Tashkent, Uzbekistan
amirkulov.edu01@gmail.com

Sabohat Kenjayeva Eshmamatovna
Karshi State University
Karshi, Uzbekistan
nilu.75@mail.ru

Jamshid Elov Bekmurodovich
Tashkent Information Technology University
Tashkent, Uzbekistan
elov.jamshid@gmail.com

***Abstract* - This paper is a conceptual study that explores the fundamental stages of creating an Uzbek-English parallel corpus, with special emphasis on the linguistic and methodological principles of selecting the base texts. The study identifies and reviews criteria for the inclusion of texts, such as genre diversity, representativeness, alignment accuracy, and linguistic relevance. Particular attention is given to balancing modern and classical texts, as well as to the role of technological tools in achieving consistent sentence-level alignment. The pipeline and recommendations presented in this paper are based on a synthesis of existing research and are proposed as a guideline for corpus developers aiming to construct a reliable and research-oriented bilingual resource.**

Keywords: Uzbek-English corpus, text alignment, linguistic base, representativeness, bilingual corpus, corpus linguistics, parallel corpus, text selection

I. INTRODUCTION

In recent years, working with parallel corpora has gained widespread importance in the field of computational linguistics. Parallel corpora consist of sentence-aligned collections of texts written in two or more languages and are widely used in translation theory, machine translation, lexicography, language learning, and linguistic research [1]. In particular, the creation of Uzbek-English parallel corpora has become a pressing issue for elevating the status of the Uzbek language in the international academic arena and for achieving parity with English-language resources. Parallel corpora have played a decisive role in the development of MT models—from phrase- and sentence-based Statistical Machine Translation (SMT) [2] to modern Neural Machine Translation (NMT) approaches [3].

The process of creating a parallel corpus consists of several key stages: text selection, cleaning and standardization, segmentation, alignment, and lemmatization. Each of these stages requires specific approaches and technical tools [4]. When selecting linguistic base, criteria such as representativeness, stylistic diversity, contemporaneity, and actual usage frequency become critically important [5]. The pipeline described in this paper is conceptual and based on a synthesis of prior research and best practices; it has not yet been implemented as a working system. The aim is to provide a comprehensive guideline for future corpus developers. The provided suggestion may really be helpful in future directions related to Uzbek-English parallel corpus giving and outlaying some basic steps to build a high-quality corpus.

II. ARCHITECTURAL PIPELINE SCHEME FOR CREATING AN UZBEK-ENGLISH PARALLEL CORPUS

It is important to conceptualize the process of creating a parallel corpus as a step-by-step and modular pipeline. Below, we describe such an architectural pipeline in stages:

A. *Data collection (Text gathering)*

In the first stage, texts are collected from selected source types. This may involve digitizing documents (if they exist in paper form, using OCR), web scraping from websites, or gathering ready-made electronic files. As a result, a collection of Uzbek and English texts is formed separately. The outcome of this stage is an unstructured collection consisting of N Uzbek texts and their corresponding English translations.

B. *Data cleaning*

In this stage, the cleaning operations mentioned earlier are carried out: converting encoding to UTF-8, removing unnecessary characters, and splitting the text into sentences. As a result, the text from each source becomes a sequence of segments. Then, each Uzbek segment is paired with its corresponding English segment (alignment), or if not done yet, this is resolved in the next alignment stage. Additionally, during cleaning, software may insert identifiers indicating which lines correspond to each other (e.g., line numbers).

C. *Extralinguistic annotation (metadata tagging)*

In this optional step, extralinguistic metadata is attached to each text (or even each segment) using XML tags. For example: text title, author, year of creation, text genre, source type (from the categories listed above), and original language (which is the source and which is the translation). This facilitates filtered analysis of the corpus later (e.g., queries like "Determine word frequency in literary works only"). Such metadata is represented using XML tags, as in this example:
<text id="0001" genre="literary" source-lang="uz" target-

lang="en" author="Abdulla Qodiriy" year="1926">. This step is optional and depends on the corpus compiler's judgment and available data.

D. Sentence alignment

Although aligners are used in various areas of NLP, their main task is to match text segments given in the source language(SL) to text segments in the target language(TL) [8]. This is one of the core stages—aligning segmented parallel texts sentence by sentence. If the files have already been paired, this stage involves comparing sentence order within each file to identify 1-1, 1-2, 2-1, or 2-2 alignments. At this stage, the heuristic, statistical, or hybrid alignment algorithms discussed earlier may be used. In most real-world projects, the following is done: first, fast heuristic alignment is applied based on punctuation and length, then results are refined using a dictionary-enhanced tool like HunAlign (statistical + heuristic), and finally, problematic cases are flagged for manual review. A human editor may intervene to filter out incorrect alignments, mark missing translations with <empty>, etc. This significantly improves quality, but takes time. Therefore, for very large corpora, validation may be done on a sample portion or via crowdsourcing. After validation, a corpus remains that contains only correctly aligned and accurately translated sentence pairs.

E. Tokenization, lemmatization, POS-tagging (linguistic processing): The next step is to annotate both sides of the parallel corpus linguistically. Although optional, this is very useful for research purposes.

- Tokenization refers to splitting sentences into words (often already done during cleaning, as aligners typically operate on tokens).

- Lemmatization converts each word to its dictionary form (e.g., *boringiz* → *bor*, *went* → *go*).

- POS-tagging assigns grammatical categories to words (e.g., noun, verb, adjective). These processes are performed separately for the Uzbek and English parts. For English, standard POS-taggers from NLTK or spaCy can be used. For Uzbek, models introduced by B. Elov or services from the Uzbek National Corpus can be used. The website uznatcorpora.uz provides automatic tagging for large texts. If online services are unavailable, localized software models trained on Uzbek need to be used. Currently, both foreign and local researchers have developed trained POS-taggers for Uzbek. At the end of this step, every word in the parallel corpus is annotated with its lemma and grammatical category. For instance, this can be represented in an XML-based system as follows (for illustrative purposes):

```
  <tu id="12345">
    <tuv xml:lang="uz"><seg>Men<tok lemma="men"
pos="PRON"> men</tok>
          <tok lemma="kitob"
pos="NOUN">kitobni</tok>
          <tok lemma="oʻqi"
pos="VERB">oʻqidim</tok>.</seg></tuv>
    <tuv xml:lang="en"><seg><tok lemma="I"
pos="PRON">I</tok>
          <tok lemma="read"
pos="VERB">read</tok>
          <tok lemma="the" pos="DET">the</tok>
          <tok lemma="book"
pos="NOUN">book</tok>.</seg></tuv>
  </tu>
```

Inside each 'tu' *translate unit*, every word is enclosed with a token tag and contains lemma and POS attributes. Of course, the actual format of a real corpus may differ, but in principle, the result is an enriched parallel corpus of this kind.

6. Word-level alignment: This stage identifies the smallest detail—which word in the parallel sentence corresponds to which translated word. This can be done automatically using GIZA++ or FastAlign (and sometimes it is handled simultaneously with Step 5). GIZA++, developed at the University of Aachen in 1999, is a statistical machine translation toolkit that includes word matching tools between parallel corpora [9]. Fast_align, created in 2014, is an open source word alignment tool[10]. The result of word alignment may include mappings like: *men → I*, *kitobni → book*, *oʻqidim → read*. This information can also be embedded into the corpus using <align> tags or via ID references on each token. Word alignment allows for deeper analysis of the translation process (for example, when one Uzbek word translates into two English words, alignment shows which part corresponds to which word—or helps identify unnecessary words). GIZA++, for instance, performs alignment in both directions and merges them using the grow-diag-final algorithm to produce a set of pairs. After this stage, the parallel corpus can be considered fully constructed.

7. Analysis and comparative statistics: Once the corpus is complete, it can be used for scientific purposes. For example, generating various statistics on the alignment between Uzbek and English: most frequently translated words, rarely translated words, elements omitted in translation, or elements that were added unnecessarily. A programmatic approach is essential here as well. In Python, such analysis can be done using pandas DataFrames, or aggregate queries can be run directly on the database. As corpus size increases, manual analysis becomes impractical, making programmatic analysis essential.

It is important to emphasize that although the steps above have been presented as a linear sequence, in practice they may be iterative. For example, sentences may be aligned, and during validation, some errors are corrected, which leads to returning to the cleaning step to update segmentation, and so on. Therefore, to manage the corpus creation process effectively, the pipeline is often implemented as a series of scripts. For instance, a makefile or snakemake workflow can be set up to call several Python scripts in sequence.

According to the proposed algorithm, an ideal parallel corpus construction process should include the following stages: text collection, cleaning, extralinguistic annotation, sentence alignment, validation, tokenization, lemmatization, POS tagging, and word alignment. We have explained this exact sequence above. Thus, this set of processes constitutes the normative architectural pipeline for creating a parallel corpus and is also recommended in scholarly literature. The architectural pipeline can be simplified and expressed in the following block diagram format and it can be clearly visualized via Fig. 1:

Text collection → Cleaning/Segmentation → Pairing → Sentence alignment → Manual validation → Tokenization → Lemmatization → POS tagging → Word alignment → Finalized parallel corpus

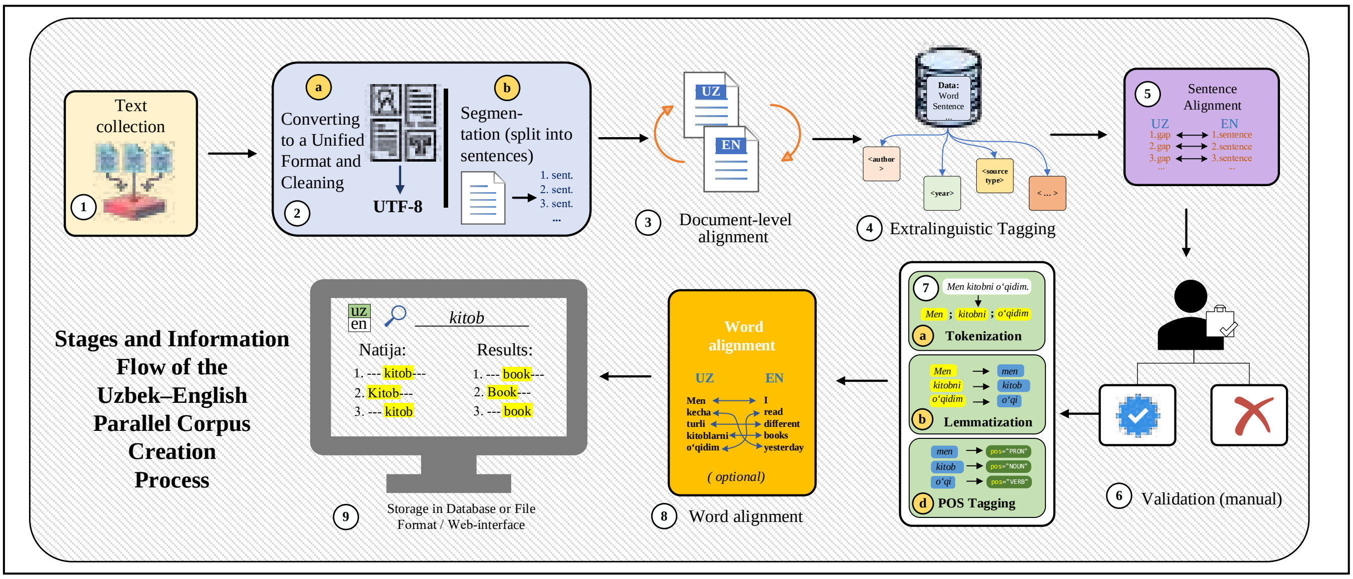

Fig. 1. Stages and Information Flow in the Creation of the Uzbek-English Parallel Corpus

In this pipeline, the output of each intermediate stage becomes the input for the next. For example, cleaned and segmented texts proceed to the pairing stage; aligned sentences proceed to the tokenization stage, and so on.

The computational linguistics approach shows that the process is automated as much as possible. In particular, tokenization, lemmatization, and POS-tagging are now highly automated. Even in the sentence alignment stage, automatic tools often provide sufficiently accurate results, requiring only quality control. Thanks to such automation, large-scale parallel corpora are now emerging. For instance, the English–Arabic parallel corpus created from World Bank or UN documents contains over 3 million sentence pairs. This is an enormous volume, achievable only through computer-assisted processing.

The Uzbek–English parallel corpus is still in its early stages, but with the use of the principles and tools described above, it is feasible to build a multi-million-word collection in the near future. This would be a major advancement not only in linguistic theory but also in applied and computational linguistics, as it would lay the groundwork for research and development in areas such as machine translation, contrastive linguistics, and corpus-based linguistic analysis.

III. Principles for Selecting a Linguistic Base for the Uzbek-English Parallel Corpus

The linguistic base of a parallel corpus refers to the composition and quality of the collection of texts included in the corpus. When forming such a base, the selected texts must encompass a variety of genres and styles, syntactic structures, thematic domains, as well as different stylistic and morphological levels [6]. The goal is to ensure the representativeness of the corpus—that is, to make the corpus a "mirror reflecting all the possibilities of the language." Representativeness is one of the most essential factors in corpus construction and holds great importance both theoretically and practically. To ensure representativeness, materials that cover diverse genres and styles are selected, enabling comprehensive study of the language's features [7].

When planning the corpus composition, the set of selected texts must provide the broadest possible representation of the language. As demonstrated by early corpus projects such as the Brown Corpus, the broader the coverage of a corpus, the more completely it can reflect the language. For example, the Brown Corpus included 500 texts from various genres, aiming to create a representative sample of English. Similarly, for the Uzbek–English parallel corpus, a balance of genres and styles must be maintained. The concept of representativeness plays a crucial role here and is explained by the principle that "a corpus is a mirror that reflects all possibilities of a language, and the clarity of that mirror is its degree of representativeness."

Therefore, when selecting the content of the linguistic base, the aim should be to fill the corpus with materials that cover the widest possible range of the language, representing the general language system rather than a narrowly defined subset.

Based on the above, we now outline the main requirements for the genre-stylistic coverage of the linguistic base in a parallel corpus using Fig. 2:

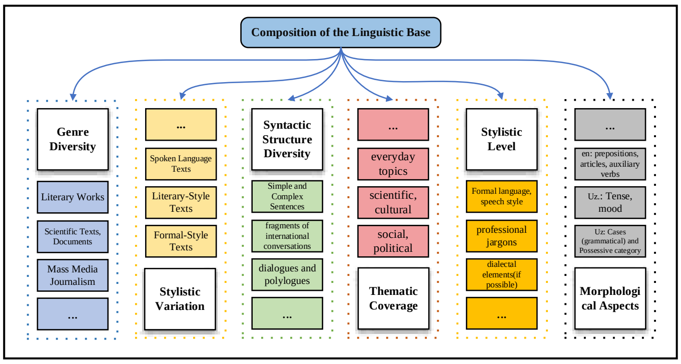

Fig. 2. Composition of the Linguistic Base

1. Genre diversity: Texts from various genres such as fiction, academic writing, official documents, mass media journalism, technical manuals, and so on.
2. Stylistic variation: Texts in formal and literary styles, texts containing elements of spoken language (e.g., interviews), and others.
3. Syntactic structure diversity: Inclusion of simple and complex sentences, dialogues and polylogues, and even fragments of international conversations.
4. Thematic coverage: Texts on a wide range of topics (social, political, scientific, cultural, and everyday themes).
5. Stylistic level: Representation of formal language, conversational style, professional jargon, and dialectal elements (if available) within the corpus.
6. Morphological features: The corpus should contain all major morphological categories of the Uzbek language (e.g., for nouns – case and possession; for verbs – tense, mood, etc.) and their equivalents in English (through prepositions, articles, auxiliary verbs, etc.).

By adhering to these principles, the resulting parallel corpus will be multifaceted and balanced, fully reflecting the lexical, grammatical, and stylistic richness of the language. Consequently, such a corpus will serve as a highly valuable resource for both linguistic research and machine translation.

IV. Main Source Types and Criteria for their Selection

Obtaining texts for a parallel corpus is one of the most critical stages. If the source types are not chosen correctly, the corpus may have limited coverage or contain low-quality translations. For the Uzbek–English parallel corpus, the following primary source types should be considered:

1) Legal documents: Laws, resolutions, government documents, international treaties, constitutional texts, and similar materials. These types of documents are typically written in formal style and are often officially translated from Uzbek to English or vice versa. For instance, the Constitution of the Republic of Uzbekistan, laws adopted by the Uzbek government, and international agreements usually have official translations. Legal texts are particularly suitable for parallel corpora due to their consistent and clear terminology, simple and formal sentence structures, and the availability of bilingual versions. Their translation quality is generally high since they are approved by official authorities.

2) Official website translations: This includes bilingual content from government agencies, international organizations, or large companies—such as announcements, news, and press releases published in both Uzbek and English on their official websites. For example, the websites of the Ministry of Foreign Affairs of Uzbekistan, the Presidential Press Service, and various ministries often provide information in both languages. Although these texts are usually brief, they are current and written in a formal-publicistic style, allowing researchers to analyze equivalence in formal registers. However, it should be noted that the English sections of some websites may not offer full translations of the original Uzbek texts—they might be shortened or summarized instead.

3) Literary works: Classic works of Uzbek literature and their English translations are especially valuable for parallel corpora. Literary texts are rich in language, stylistically diverse, and full of metaphors, expressions, and idioms. Much of the beauty of the Uzbek language is best showcased through literature, which is why including high-quality literary translations in early stages of corpus development is highly recommended. The main challenge, however, is that literary translations are relatively rare—only certain classic novels and short stories have been translated. Examples include works by Abdulla Qahhor, Cho‘lpon, and O‘tkir Hoshimov (e.g., *O‘tkan kunlar*, *Kecha va kunduz*, *Dunyoning ishlari*), which are recommended for inclusion. These are professionally translated and offer invaluable material for comparative analysis of expressive and artistic language.

4) Mass media texts (news media): Newspapers, magazines, and online news websites that publish content in both Uzbek and English can also serve as sources. Two scenarios are possible: some international news agencies translate their content into Uzbek, or domestic sites publish content in foreign languages (e.g., kun.uz occasionally posts news in English). The style of mass media texts is typically publicistic, written in clear and accessible language, and reflects modern linguistic trends and neologisms. Some media texts may also include informal expressions. Collecting parallel text pairs from news media is relatively easier, as the texts are short and follow standard formats.

These four categories represent the primary sources for a parallel corpus. The selection criteria dictate that sources must be reliable and high quality, since each segment (sentence) in the corpus must have an accurate translation. Therefore, sources with professional translations are preferred: official documents from government agencies, published translated books from established publishers, press releases from reputable news agencies, etc. When selecting sources, it is crucial to find “source text – translated text” pairs whenever possible. If a document or literary work exists in both original and translated form, it is ideal for inclusion in the corpus.

It should be especially emphasized that matching sources across genres and types is not easy. Finding translations of different genres from one language into another is a challenging task [7]. For instance, while journalistic and official texts are frequently translated between English and Uzbek (particularly in the field of government communication), literary translations are quite limited. Therefore, when building the parallel corpus, authors have chosen to include as many high-quality literary translations as possible in the initial stage—to showcase the beauty of the language. Additionally, when a sufficient number of official and media text translations are found, they should also be incorporated.

The principles for selecting linguistic base sources can be summarized as follows:

1. The texts must be high-quality translations (preferably professionally translated);

2. The texts should be diverse in genre and style (covering the various categories mentioned above);

3. The topics of the texts should vary (this allows the corpus to reflect language usage in different contexts);

4. The length and structure of the texts should be appropriate (extremely short or overly long texts can cause difficulties in processing, so texts with average-length segments are ideal);

5. ***The text pairs must be clearly matched*** *(i.e., the original and its full translation must be available—not just partial or loosely interpreted versions).*

Searching for and selecting sources that meet these criteria is the initial and decisive step in building a parallel corpus. To compare the types of sources mentioned above and their characteristics, we have compiled them into the following Table I.

TABLE I. Types of Parallel Corpus Sources and Their Characteristics

Source Type	**Examples (from works)**	**Characteristics (translation quality and style)**
Legal documents	*Constitution, laws, government resolutions*	*Formal-style texts, clear terminology. Official translations available, high quality and well-aligned.*
Official websites	*Government and agency announcements, press releases*	*Formal style, brief information. Versions in both languages often available, but content may be limited in volume.*
Literary works	*O‘tkan kunlar (Abdulla Qodiriy) and its translation; Kecha va Kunduz (Cho‘lpon); Dunyoning ishlari (O. Hoshimov), etc.*	*High literary style, rich vocabulary and expressions. Translation quality is high (professional). Rare, but shows linguistic elegance and expressiveness.*
Mass media texts	*BBC Uzbek ↔ BBC English news reports; Kun.uz news (Uz/En)*	*Publicistic style, simple language. Translations available but not always. Contains modern language elements and jargon. Translation quality needs manual verification (may misalign).*

As seen in **Table 1** above, each type of source has its own unique value and contributes to the parallel corpus in a distinct way. Ideally, the corpus should include **all of these source types**, thereby becoming a **multi-genre parallel corpus**. Only in this way can the corpus provide the necessary material—for instance, for translating legal texts or for analyzing literary texts—ensuring that suitable examples can be found for each use case.

V.Expected Future Results

The development of the Uzbek-English parallel corpus following the proposed pipeline is expected to yield significant benefits for both linguistic research and practical applications. Upon completion, the corpus will serve as a comprehensive, balanced, and representative resource reflecting the lexical, grammatical, and stylistic richness of both languages. In particular, the following outcomes are anticipated:

- Creation of a high-quality bilingual corpus containing several million sentence pairs, aligned at the sentence and word levels, and enriched with linguistic annotations such as lemmatization and part-of-speech tags.

- Availability of a public resource for researchers in the fields of corpus linguistics, translation studies, and language education, fostering further study of cross-linguistic phenomena.

- Contribution to the development of more accurate and contextually aware machine translation (MT) systems for Uzbek-English and vice versa, particularly benefiting neural MT models that rely on large, clean parallel datasets.

- Support for building domain-specific terminological databases and glossaries, assisting translators, lexicographers, and language technologists.

- Enhancement of the visibility and internationalization of the Uzbek language by providing a standardized, accessible linguistic resource that meets global research standards.

- Opportunities for future extension of the corpus to additional language pairs involving Uzbek (e.g., Uzbek-Russian, Uzbek-Turkish), thereby establishing a foundation for multilingual corpus initiatives.

By adhering to the principles outlined in this study, the resulting corpus is expected to advance both theoretical and applied research in computational linguistics, strengthen the position of the Uzbek language in digital and academic domains, and inspire further collaborative projects in the region.

VI.Conclusion

The process of creating an **Uzbek–English parallel corpus** is a complex and multi-stage endeavor that requires not only technical procedures but also a deep **linguistic approach**. This article has thoroughly examined the main stages of corpus construction—**data collection**, **cleaning**,

alignment, and **annotation**. In particular, the **linguistic base** and the **principles guiding its selection** play a critical role, as they determine the corpus's level of **representativeness** and **balance**. When selecting materials, it is essential to consider the **functional purpose** of the corpus, **genetic and functional styles**, types of texts, and their **genre classification**. Furthermore, it is recommended that materials be selected only from **authentic natural language samples** that are well-founded and properly documented. .

REFERENCES

[1] I. Simeon, "Parallel corpora and multilingual dictionaries," *Filologija*, no. 38–39, pp. 0–0, 2002. [Online]. Available: https://hrcak.srce.hr/173311

[2] D. Banik, A. Ekbal, and S. C. Satapathy, "Fuzzy influenced process to generate comparable to parallel corpora," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, just accepted, June 2023. https://doi.org/10.1145/3599235

[3] https://quod.lib.umich.edu/cgi/t/tei/tei-dx?type=pointer&value=SACSAL

[4] F. Forgac, D. Munkova, M. Munk, and L. Kelebercova, "Evaluating automatic sentence alignment approaches on English-Slovak sentences," *Scientific Reports*, vol. 13, no. 1, p. 20123, 2023.

[5] S. Paun, "Parallel text alignment and monolingual parallel corpus creation from philosophical texts for text simplification," in *Proc. 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, June 2021, pp. 40–46.

[6] F. Zanettin, "Parallel corpora in translation studies: Issues in corpus design and analysis," in *Intercultural Faultlines*, Routledge, 2017, pp. 105–118.

[7] A. Eisele and Y. Chen, "MultiUN: A multilingual corpus from United Nation documents," in *LREC*, May 2010.

[8] N. Matyakubova, A. Dauletov, Sh. Khamroyeva, B. Mengliyev, and E. Adali, "Algorithm of creating the 'Uzbek-English aligner' program," in *Proc. 8th Int. Conf. on Computer Science and Engineering (UBMK)*, Mehmet Akif Ersoy University, Burdur, 2023.

[9] F. Och and H. Ney, "Improved statistical alignment models," in *Proc. 38th Annu. Meeting of the Association for Computational Linguistics*, Hong Kong, China, Oct. 2000, pp. 440–447.

[10] Ch. Dyer, V. Chahuneau, and N. Smith, "A simple, fast, and effective reparameterization of IBM Model 2," in *Proc. North American Chapter of the Association for Computational Linguistics*, June 2013.