AJLLSE GLOBAL RESEARCH NETWORK

AMERICAN Journal of Language, Literacy and Learning in STEM Education
Volume 02, Issue 12, 2024 ISSN (E): 2993-2769

Uzbek Dialect Corpus Database

Abdurahmanova Sayyora
Doctoral candidate at the Tashkent State University of Uzbek Language and Literature named after Alisher Navoi

Abstract. *This article discusses the linguistic foundations of the corpus of shevas and the principles of compiling a database of the corpus of Uzbek dialects. Information on the stages of compiling a corpus of data, the structure and size of the data warehouse is given. At the same time, it is thought what should be the emphasis on the formation of the database, what technical and linguistic work should be carried out in case of launching the dialectal corpus.*

Key words: *frequency analysis, correlation, metadata addition, metadata, lexical annotation, syntactic annotation, semantic annotation, sentiment annotation.*

Every literary language is a refined and polished version of a national language, standardized and regulated. It can be defined as follows: a literary language is a form of the national language that is lexically stable, phonologically and grammatically structured, and adheres to orthographic and orthoepic norms. Each language has its foundational dialect, which serves as the basis for the literary language and acts as a source for its enrichment.

Although the formation of the Uzbek literary language has relied to some extent on all dialects, certain dialects have played a fundamental role as the lexical and phonetic base of the literary language. “Indeed, the formation of the Uzbek literary language involves all dialects to some degree, but specific dialects serve as the foundation or pillar for the literary language. That is, the Uzbek literary language adopts lexical, phonetic, and grammatical features from certain dialects or groups of dialects as facts of the literary language and develops alongside the evolution of these dialects or dialect groups. Since dialects are a living form of language, they are constantly evolving and changing. Consequently, the literary language advances in connection with the development of its foundational dialects. Conversely, if a literary language lacks a foundational dialect, it will gradually fall out of use,” [Zulfiya Xolmirzayevna Rustamova https://doi.org/10.5281/zenodo.7423965].

Dialects, which serve as the living language among the people, are a manifestation of sociolinguistics. Preserving them today holds significant importance for the historical development of our language. However, urbanization and the decline in the number of dialect speakers bring forth the task of creating a dialectal corpus. The first step is to establish a database for the corpus. It is crucial to address the questions: What is a database, and what is its purpose?

A database is not merely a collection. It is a unified compilation structured according to specific requirements and governed by predetermined commands within the scope of those requirements.

A central repository for data, known as a data warehouse, is a type of database designed specifically for quick queries and analysis. Databases can be classified as open, closed, static, or continuously updated. The primary function of a database is to elucidate the functionality, operating mechanism, and convenience of a specific system. The completeness of a database ensures the efficient operation of a given system or program.

According to global standards, databases designated for corpora are characterized as follows: First, the purpose of creating the corpus must be determined. This purpose defines the field of application. It is essential to identify which aspects of the language—such as morphology, syntax, semantics, pragmatics, or lexicon—the corpus will analyze. This purpose serves as a guiding factor in forming the corpus and organizing the database.

The next step involves determining the type of corpus, which focuses on selecting the type of material (text, speech, audio, video) to be collected and deciding what information needs to be extracted from these materials. Following this, the process of data collection or source selection begins. Sources may include scientific literature, books, articles, materials from mass media, text from internet forums, or audio and video recordings.

Once the data is gathered, the size of the corpus is determined according to its purpose. The larger the corpus, the broader its analytical possibilities; however, this may complicate the size of the database and its processing stages.

In constructing and formatting the corpus, key stages include formatting the text and adding metadata. Text formatting involves standardizing all texts in the corpus. This includes cleaning the text (e.g., removing extra spaces), formatting (e.g., in HTML, XML, JSON), and normalizing it to facilitate later analysis. Adding metadata entails appending information for each element (text, word, or sentence) in the corpus. Examples of metadata include the word category, the source of the text, the time, or the location of its creation.

Annotating data is also an essential step in this process and can be categorized as follows:

1. **Lexical Annotation**: Annotating each word based on its meaning, grammatical category, and syntactic role. For example, specifying the word type (verb, noun, adjective, etc.).
2. **Syntactic Annotation**: Identifying sentences in the text and their structures. Using syntax analysis to determine the grammatical structure of a sentence.
3. **Semantic Annotation**: Adding semantic meaning to the data. This is particularly important for analyzing data from different perspectives. **Sentiment Annotation**: Identifying and labeling positive, negative, or neutral emotions in the text.

After completing the stages mentioned above, the corpus must be classified and analyzed. Collected data should be categorized based on topics, languages, dialects, styles, or other parameters. If the corpus includes multiple languages, each text should be classified by language. Analyzing the data within the corpus involves computational linguistics methods, such as Natural Language Processing (NLP) technologies. Since these technologies may not fully accommodate every language, it is crucial not to overlook the human factor in this process. For languages like Uzbek, it is currently challenging to completely standardize or adapt them to specific programs.

In the next stage, organizing the structure of the database becomes essential. Attention must be given to the database system, tables and structure, and indexing. Either relational (SQL) or non-relational (NoSQL) systems can be chosen to create the database.

- **SQL systems** (e.g., MySQL, PostgreSQL) store data in table form.
- **NoSQL systems** (e.g., MongoDB, Elasticsearch) are more suitable for handling large volumes of data.
- Tables and structures involve storing each text, word, sentence, or phrase in separate tables within the corpus. Additionally, metadata, annotations, and other attributes linked to each element should also be stored.
- **Indexing**: Indexing is necessary for quickly searching the data. For example, creating indexes for words and phrases enables faster search and analysis processes.

In the **Interface and Search System** stage, a user interface and search system are developed. This interface allows users to search for words, phrases, and lexical combinations within the corpus. The search system facilitates searching texts, displaying annotations, and filtering data.

The **Export and Update** stage involves enabling the export of data from the corpus (e.g., in CSV, XML, or JSON formats). This is valuable for scientific research and other analyses. The corpus should also be regularly updated by adding new materials, correcting errors, and introducing new annotations.

Storage and Security are critical aspects of this process. Regular backups of the corpus help prevent data loss due to technical issues. Restricting access to the corpus and ensuring confidentiality are essential for maintaining security. If necessary, additional measures should be taken to secure the database.

Collaboration with Users and Partners is also important during the creation and development of the corpus. Collaborating with other researchers or linguists helps address shortcomings and ensure the corpus's quality. This collaboration can include adding new data, correcting errors, or discussing the analysis processes within the corpus.

In this way, forming a database becomes useful for various linguistic and computational linguistics studies, enabling a deeper understanding of language.

The linguistic foundation of corpus creation refers to the set of methodological and technical tools used in linguistics to study language and its structures. Corpora represent the cornerstone of linguistics by facilitating the collection and systematic storage of real speech samples. They demonstrate their effectiveness in empirically studying language.

The linguistic foundation of corpus creation is based on the following principles:

1. **Purpose and Objectives of the Corpus**: The first step in corpus creation is defining its purpose. For instance, a corpus may be used in fields such as linguistics, sociolinguistics, linguistic typology, grammar, semantics, or stylistics. These objectives serve as the primary factors determining the structure and composition of the corpus.
2. **Systematic Collection**: The collection of linguistic data for the corpus must be systematic and standardized. The collected data should not be random but selected according to the purpose. During corpus creation, materials may be gathered from specific periods, fields, or stylistic contexts. Particular attention should be paid to the construct as a whole, formed from the sequential relationship of its meaningful parts [Z. Kholmanova 2019:229].
3. **Representativeness** [Zakharov V., Mengliyev B., Hamroyeva Sh., 2023:185]: The linguistic content of the corpus, including its linguistic resources (words, phrases, grammatical structures, phonetic features), should reflect real language data as much as possible. For instance, when creating a dialectal corpus, audio and video materials are of critical importance. The issue of representativeness in a corpus is addressed by ensuring the sufficiency and diversity of texts. According to V.P. Zakharov and S.Y. Bogdanova, special attention should be paid to determining what unit will be considered a corpus text when examining the genre-thematic structure of the corpus [Захаров В.П., БогдановаС.Ю. 2011:36.]. Representativeness ensures the reliability and scientific value of the corpus.
4. **Annotation**: Corpora are often annotated based on lexical, grammatical, and semantic features. For example, this may include identifying word types (noun, verb, adjective, etc.), syntactic structures, or explaining semantic contexts. Annotation allows for systematic study of various layers of language.
5. **Correspondence**: Classification of linguistic materials and their transfer or translation into different language corpora. This is particularly important for linguistic research, as corpora are often bilingual or multilingual to study lexical and grammatical structures of other languages. In a dialectal corpus, correspondence is crucial in linguistics for understanding the changes occurring in a language by identifying similarities and differences between dialects. A dialectal corpus contains various dialects of a particular language and presents their distinctive features. Correspondence in a corpus focuses on the following aspects:

- ✓ Identifying dialectal differences,
- ✓ Observing changes and development in language,
- ✓ Analyzing short-term and long-term language contacts,
- ✓ Examining connections between normative and local dialects,
- ✓ Accounting for dialectal differences in translation.

Overall, correspondence in a dialectal corpus plays a vital role in linguistics for understanding relationships between dialects, tracking language changes, and analyzing language development. It also aids in language teaching, translation, and standardization processes.

6. **Tools for Corpus Analysis:**

   Special software (corpus analyzers) is used to conduct linguistic analysis with corpora. These tools enable searching for words and phrases, frequency analysis (frequency distribution, normalization, chi-square test) [Frequency distribution, normalization, chi-square test], constructing syntactic diagrams, lexical analysis, and more.

7. **Form and Size of the Corpus:**

   Corpora can be selected based on their form (electronic, textual, audio) and size. They may consist of one or several million words. The size of a corpus varies depending on the language or linguistic field being studied.

8. Corpora are used in scientific research across various languages and have become a fundamental tool for practical language study. Today, they are widely applied in fields such as machine translation, speech recognition, lexicography, and natural language processing.

9. In the process of developing a database for a dialectal corpus, paying attention to the following aspects can ensure the corpus is enriched with texts. Units in a dialectal corpus allow for the analysis of elements found in folk songs, proverbs, idioms, and riddles without separating them from their context. Each unit is entered into the database alphabetically and contributes to the enrichment of the electronic dialect dictionary. This process organizes the data systematically within a short period.

10. The collected data is analyzed and edited using various methods. Initially, semantically similar words and phrases are identified and analyzed. Words that have undergone phonetic changes are also considered during this process. The recorded results are analyzed on a regional scale, and this process can lead to the creation of an electronic atlas of Uzbek dialects. Additionally, based on the included folklore samples, it is possible to identify lacunar units or gain a clearer understanding of their contextual meanings.

11. The databases prepared for corpora differ depending on their purpose and context of use. Below are the main types of databases:

1. **Corrective Corpus** (Annotated Corpus) [Developing Linguistic Corpora: a Guide to Good Practice]

Description: This type of corpus includes texts that are tagged with various lexical, morphological, or syntactic markers. These tags help identify parts of speech, syntactic structures, or semantic meanings.

Example: The English "Penn Treeban" [Penn Treebank Dataset | Papers With Code] corpus, where words are annotated with morphological and syntactic tags.

Usage: Used in automated speech analysis, translation systems, and the development of machine learning models.

2. **Categorized Corpus** [NLP | Categorized Text Corpus - GeeksforGeeks]

Description: This type of corpus consists of texts divided into various classes or categories. Each text may belong to a specific category, such as scientific, literary, social, etc.

Example: A collection of texts, such as social media posts, categorized by topic or domain.

Usage: Utilized for topic-based searches, document classification, and sentiment analysis (e.g., positive, negative, or neutral sentiments).

3. **Parallel Corpus** [Parallel Corpora | CLARIN ERIC]

Description: This type of corpus contains translations of the same text in two or more languages, providing parallel texts for each language.

Example: Official documents of the United Nations or international organizations available in multiple languages (e.g., English and French).

Usage: Used in machine translation (MT), language learning systems, and bilingual analysis.

4. **Bag-of-Words (BoW)** [Bag of words (BoW) model in NLP - GeeksforGeeks] **Corpus**

Description: In this corpus, texts are represented as separate word vectors, counting the frequency of each word but ignoring word order and grammatical structure.

Example: Information about the words in a text and how often they are used.

Usage: Used in text classification, sentiment analysis, and other machine learning tasks.

5. **Tokenized Corpus** [python - How to tokenize a text corpus? - Stack Overflow]

Description: In this corpus, texts are broken down into words or tokens, with each token treated as an individual unit.

Example: The sentence "I go to school" could be tokenized as: ["I", "go", "to", "school"].

Usage: Used for word analysis, meaning identification, morphological, and syntactic analysis.

6. **Sentiment Corpus** [BeSt: The Belief and Sentiment Corpus - ACL Anthology]

Description: This corpus consists of texts categorized based on positive, negative, or neutral sentiments.

Example: Social media posts or user reviews categorized by sentiment (e.g., positive or negative).

Usage: Used in sentiment analysis, studying user opinions, and analyzing brand or product feedback.

7. **Temporal Corpus**

Description: Texts in this corpus are classified chronologically, reflecting specific periods or changes over time. [Neuroanatomy, Temporal Lobe - StatPearls - NCBI Bookshelf].

Example: News articles or blog posts linked to specific dates.

Usage: Used for analyzing texts corresponding to temporal changes, tracking trends, and conducting historical analyses.

8. **N-gram Corpus** [N-grams: based on one billion word COCA corpus]

Description: This corpus represents texts as sequences of words (n-grams). For example, a 2-gram (bigram) includes pairs of words like: "I go", "go to", "to school".

Example: Analyzing word sequences and building language models.

Usage: Used in language modeling, speech recognition, machine translation, and language learning systems.

9. **Diachronic Corpus** [Diachronic Corpora | SpringerLink]

Description: A diachronic corpus includes texts based on historical or temporal variations, tracking changes in language over time.

Example: Historical texts, classic literature, and words or expressions analyzed for their evolution over time.

Usage: Used in historical linguistic analysis, studying language development, and analyzing language evolution.

10. **Specialized Corpus** [EJ1145032.pdf], [Specialized Monolingual Corpora in Translation].

Description: A corpus consisting of texts focused on a specific domain or subject, such as medicine, economics, technology, or law.

Example: A medical corpus containing texts and terms exclusively related to the medical field.

Usage: Used for domain-specific processing, terminology studies, and specialized field analyses.

Each type of corpus serves specific purposes and usage scenarios, primarily in machine learning, natural language processing (NLP), translation systems, sentiment analysis, and various other fields.

REFERENCES

1. Rustamova Zulfiya Xolmirzayevna TDPU Boshlang‘ich ta’limda ona tili va uni o‘qitish metodikasi kafedrasi dotsenti v.b https://doi.org/10.5281/zenodo.7423965
2. Z.Xolmanova Kompyuter lingvistikasi o‘quv qo‘llanma. “Aktiv print” Toshkent 2019 229-b.
3. Zaxarov V., Mengliyev B., Hamroyeva Sh., Korpus lingvistikasi[Matn]: O‘quv qo‘llanma. Toshkent, 2023. – 185b
4. Захаров В.П., БогдановаС.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2011. – С. 36.
5. Frequency distribution, normalization, chi-square test
6. Penn Treebank Dataset | Papers With Code
7. Developing Linguistic Corpora: a Guide to Good Practice]
8. NLP | Categorized Text Corpus - GeeksforGeeks
9. Parallel Corpora | CLARIN ERIC
10. Bag of words (BoW) model in NLP - GeeksforGeeks
11. Python - How to tokenize a text corpus? - Stack Overflow
12. BeSt: The Belief and Sentiment Corpus - ACL Anthology
13. Neuroanatomy, Temporal Lobe - StatPearls - NCBI Bookshelf
14. N-grams: based on one billion word COCA corpus
15. EJ1145032.pdf
16. Specialized Monolingual Corpora in Translation