Web of Discoveries: Journal of Analysis and Inventions
webofjournals.com/index.php/3

Volume 3, Issue 5, May - 2025 **ISSN(E):** 2938-3773

LINGUISTIC ANNOTATION OF A CORPUS OF OFFICIAL STYLE TEXTS

Sanjar Amirqulov
Teacher, Denau Institute of Entrepreneurship and Pedagogy

Abstract
This article explores the issues related to the creation and linguistic annotation of a corpus of Uzbek texts in the official style that emerged during the independence period. It describes linguistic annotation models that enable automatic and semi-automatic analysis of official texts based on the theory of corpus linguistics. The article presents a methodology developed on the basis of research into morphological, syntactic, and semantic annotation types, as well as statistical results obtained from the corpus.

Keywords: Official style, text corpus, linguistic annotation, morphological analysis, corpus linguistics, independence period, annotation methodology.

Introduction

The specific features of the official style of the Uzbek language, especially after the independence period, began to manifest actively in legal, administrative, and organizational domains. In analyzing texts that emerged in these areas, linguistic corpora serve as a crucial primary source. From this perspective, the construction of a corpus of texts in the official style from the independence period and its linguistic annotation is one of the pressing tasks facing modern corpus linguistics.

During the independence period, the style of Uzbek official-administrative documents was restructured. Numerous new Uzbek-derived lexical items were introduced into its lexicon, and the morphological and syntactic norms of the official style were reconsidered. This situation has placed a responsibility on linguists to analyze the updated texts of official-administrative documents from a linguistic point of view and to identify the factors that have influenced the development of this style. [3;4]

Annotation (Linguistic Annotation) is the process of marking up a text from a linguistic point of view—that is, attaching various linguistic data to words or sentences in the text with the help of computational tools. In the literature, this is also referred to as linguistic tagging. A linguistic tag can be of different types such as morphological tags or syntactic tags. This process is considered one of the key stages in corpus linguistics and computational linguistics. On this basis, annotation is classified into several types:

1. Morphological annotation – marks the part of speech and grammatical form of a word.
2. Syntactic annotation – identifies syntactic relations between words (subject, predicate, etc.).
3. Semantic annotation – assigns meaning, theme, or role to the word.
4. Pragmatic or discourse annotation – identifies stylistic and communicative features of the text.

Example: “The Constitution of the Republic of Uzbekistan was adopted.” When this sentence is annotated, each word is assigned information such as its part of speech, morphological form, and syntactic role.

Word	Part of Speech	Morphological Form
O‘zbekiston	**Noun**	**Nominative case, singular**
Respublikasining	**Noun**	**Genitive case, possessive suffix**
Konstitutsiyasi	**Noun**	**Nominative case, possessive, singular**
qabul qilindi	**Verb**	**Past tense**

When annotating this sentence, each word is assigned information such as its part of speech, morphological form, and syntactic role.

The stylistic units characteristic of official texts from the independence period were systematically analyzed, and a specialized model of linguistic annotation was developed based on the corpus (morphological + syntactic + pragmatic).

For official texts, an annotation module was developed on the uzbekcorpus.uz platform.

Statistical indicators were analyzed across formal expressions, official abbreviations, and structural units within the texts.

The described model was evaluated as a universal analytical method that could be applied to corpora of other functional styles as well.

The annotation of texts in the official style was carried out in the following stages:

Within the scope of the research, 500 official Uzbek documents—including decrees, laws, resolutions, and regulations—were selected. Initially, these documents were available in Word format, and special conversion processes were undertaken to adapt them to the text corpus. Microsoft Word documents were converted into plain text format (TXT).

Since some of the texts included in the corpus were in HTML or PDF formats, technical codes, symbols, and unnecessary formatting elements were cleaned from these documents. Morphological annotation was applied to all words in the corpus texts from the perspectives of lexical units, parts of speech, and grammatical features. Subsequently, sentences and their structures, i.e., sentence constituents and their functional roles, were identified.

At this stage, forms of address, speech acts (commands, requests, recommendations, etc.), and the degree of formality of the documents were determined. This annotation was carried out partly automatically and partly manually. Based on this, semantic-pragmatic models were planned for development, and statistical analyses were performed using the data in the corpus.

During this process, word frequency, repetition of grammatical constructions, and other linguistic units were analyzed graphically.

The operations described above are summarized in the following table:

The average distribution of parts of speech used in the texts is as follows: nouns constitute the largest share, accounting for 38%. Verbs come second with 24%. Adjectives follow, making up 12% of the total words. Adverbs account for 6%, prepositions for 9%, and numerals, pronouns, and other parts of speech make up a total of 11%.

Syntactic structures in official documents exhibit distinctive features. According to the analyses, simplified official sentences, i.e., one-sided constructions, are the most widespread, constituting 41% of all sentences. Complex compound sentences hold the second place with 32%. Sentences complicated by auxiliary means account for 15%. Imperative sentences are relatively less frequent, comprising 12%. (For example, expressions such as “mazkur” (“the present”), “yuqoridagilardan kelib chiqib” (“based on the above”), and “O‘zbekiston Respublikasi Vazirlar Mahkamasi qarori” (“Decree of the Cabinet of Ministers of the Republic of Uzbekistan”) belong to the most frequent phrases.)

This article presents the creation of a corpus of Uzbek texts in the official style from the independence period and practical methods of linguistic annotation. The results demonstrate that the official style is rich in unique formal units, structures, and stylistic forms. The annotation methods developed for this corpus can also be applied to the analysis of other functional styles in the future.

References

1. Abduraxmonova N., Sadikova M. Korpus lingvistikasida matnlarni lingvistik annotatsiyalash tamoyillari // “Ilm-fan ta’limining rivojlanish istiqbollari” mavzusidagi ilmiy konferensiya materiallari, Toshkent, 2020-yil 27-aprel, B. 332–336.
2. Лутфуллаева Д. Э. Мустақиллик даври расмий-идоравий иш услуби тараққиёти. Монография. – Тошкент, 2020.
3. Viktor Zaxarov, B. Mengliyev, Sh. Hamrayeva. Korpus lingvistikasi: korpus tuzish va undan foydalanish. O‘quv qo‘llanma. Toshkent, 2021.
4. UzbekCorpus.uz – Elektron matnlar korpusi. https://uzbekcorpus.uz/