Research Article

Computational Linguistics

FORMAL-FUNCTIONAL MODELS OF THE UZBEK ELECTRON CORPUS

Keywords: Uzbek corpus, morphoanalyzer, metadata, parallel corpora, text analysis, corpus manager.

Nilufar Abdurakhmonova

PhD, Associate professor of National University of Uzbekistan, Tashkent, Uzbekistan.

Abstract

The paper is devoted to the structure and its linguistic annotation for building Uzbek Corpus. Linguistic annotation, metadata and corpus manager as formal-functional model of the corpus are important for usage for many purposes. The fact that the platform allows users to address language and literature issues, use it online. The Uzbek corpus based on structural and sub corpus models, which partially represented in this paper, is going on process to develop Uzbek language technology.

I. INTRODUCTION

It can be said at the present time that corpus is one is main tools for natural language processing and as research object for other cross fields to study language and speech knowledge. Particularly, corpus is important if the language is considered lack of resources of language as linguistics resource and platform. Today the Uzbek language one is most spoken language among Turkic languages in the world play important role for communication and accumulation of historical epoch language development. Hence creation corpus is necessary all aspects of development language technology. It must be admitting that there was not any electron corpus of the Uzbek language before the project achievement. This corpus is one of the first corpuses for Uzbek which created as electron version.

The Uzbek corpus as a part of the project CUFMG (It has been created and motivated under the project of “Corpus of Uzbek literary works under themes family, makhalla and equality gender” by financed the institute “Family and makhalla” scientific research institute in Uzbekistan (2020-2021, JHBL-20). Our project focused to chronological literal texts by different genres of epic works of Uzbek writers. Firstly, this corpus is web service is devoted to more investigation humanity and social problems connecting with family, makhalla and gender equality. Although collecting written Uzbek literal works as classification texts chronological aspect of abovementioned problems, it could restrict research object for Uzbek language. But later considering many aspects of corpus usage case our group decided to enhance formal-functional models of corpus with linguistic annotation.

As according to principles of corpus being corpus manager whose search system should be considered pilot formal-functional models of corpus linguistics. However, corpus differentiate with annotation (existing or not), to develop forward the next stages not only for linguistic approach but also other spheres as well.

II. REPRESENTATION OF TEXT IN UZBEK CORPUS

Obviously, there are many corpora of world languages in the world. They dedicated multifunctional aspect of language usage [1,2,3,4]. Considering experience of a number of corpora models, there are corpus managers, linguistic annotation and metadata is required for corpus to apply information for any purposes. Creation corpus is multilevel process for language processing. V. Zakharov [6] points out some requirements for contemporary corpus manager to be construct concordance, search by not only word but also collocation, search by fragments (complex query), sort list according to some requirement by user chosen, make it possible to display the found word forms in an extended context; provide statistical information on individual elements of the corpus; display lemmas, morphological characteristics of word forms and metadata (bibliographic, typological), which depends on the degree of markup of the corpus; save and print the results; work both with separate files and with corpora of unlimited size.

Each scenario of algorithm of analysis of corpus belongs to language characteristics and grammatical rules. For example, for European languages there are ready tools for morphological analysis and parsing as a stage of stemming, tokenization and lemmatization. These tools are accessible to users via WebLicht2, a tool chainer providing both infrastructural services and a GUI for combining the individual tools [7].

Encoding text of corpus is important for representation of language analysis. There are several formats to input data in corpus. A corpus is prepared by whole texts or of fragments or text samples by different genres oral or written language texts.

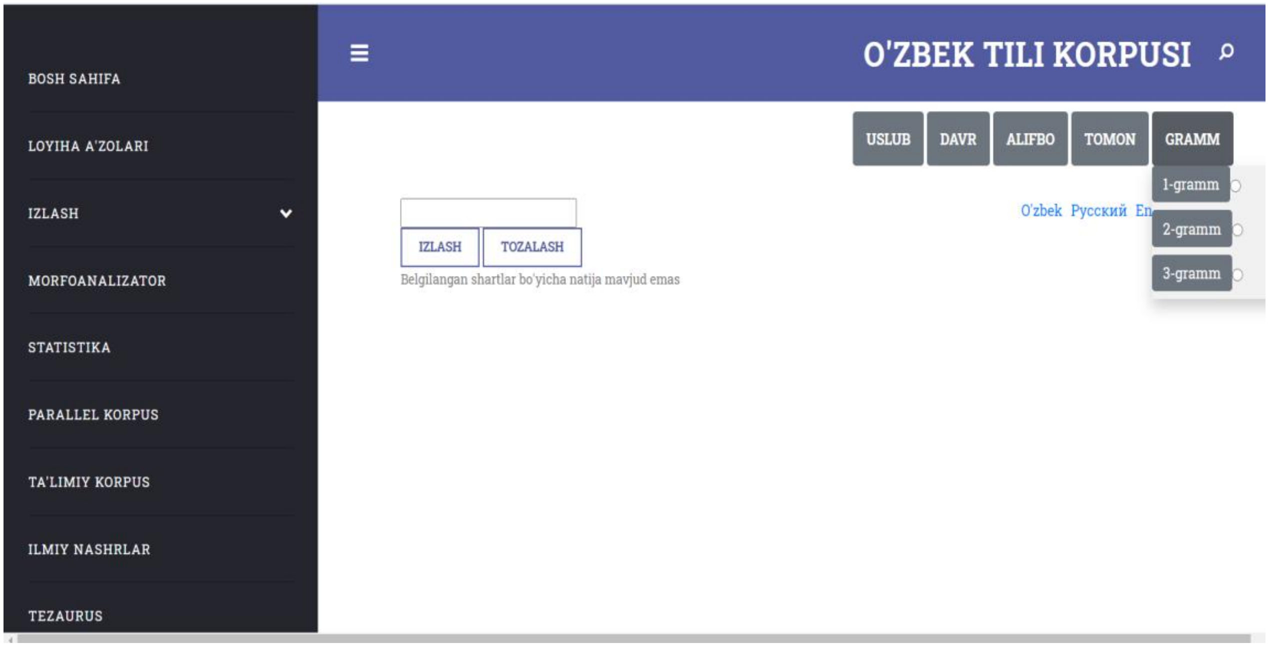
III. TEXT ANALYSIS AND CONTENT OF UZBEK CORPUS

The Uzbek language saw changeable alphabet through the years [8]:

Arabian	Latin	Cyrillic	Latin
—1929	1936—1940	1940-1992	1993

As we see there are different graphemes of Uzbek used during this time. Now Uzbek language applies two graphemes Latin and Cyrillic. Latin and Cyrillic differ orthographical rules and <***o‘*>** and <***g‘***> letters with various forms due to not existing Unicode for these letters. If words searched by Sketch engine like these words appeared division by misused words like **<*g’*> *or* <*g*`**>. Moreover, morphological features are also dislike in some occasions for both graphemes; {*tog‘+ga=>tog‘ga (Latin)} / {тоғ+га => тоққа (Cyrillic)}*. Moreover, machine readable texts should be recognized both graphemes as applied now in the society by parallel documentation and publication process. Therefore, our research comprises considering both Latin and Cyrillic graphemes should be available for users search interface.

Consequently, the conception of Uzbek corpus divided three stages time of literature according to chronological aspect of *literal* texts. It is not only for graphemes, but also political direction in this country as well. Hence they are: jaded literature (1905-1930), Soviet literature (1930-1990) and independence literature (since 1991). Makhmudxoʻja Bekhbudiy, Abdulla Avloniy, Hoji Muin, Sidqiy Hondayliq, Abdurauf Fitrat, Hamza Hakimzoda Niyoziy, Abdulla Kodiriy, Chulpon, Botu writers' works in Jaded literature input in database according to various genres: 18 stories, 3 novels, 4 dramas, 5 comedies, 1 thesis and 5 scientific articles.



Belgilangan shartlar bo'yicha natija mavjud emas

Pic.1. Corpus manager system

Parameters of the text given in metadata with the name of author, name of the work, genre, publishing house and year, Latin or Cyrillic.

IV. CORPUS MANAGER SYSTEM

Mostly, often, such systems are based on ready-made solutions, which lead to problems with the speed of search queries (data samples), system flexibility, and scalability. Our observation shows that the corpus manager of many language corpuses is controlled by ready-made technologies. These technologies provide fast content searching system. As for the national corpus of the Russian language, Yandex server is used as a tool of searching engine [8]. This search engine consists of creating direct and inverse queries, as well as logical operations when searching by logical operators and, or (conjunction, disjunction). Another is a Sketch Engine system that supports document metadata independently and uses a special, query language (CQL - Corpus Query Language) to view corpus statistics. The search engine of the case, created for the Tatar language, consists of a control panel for checking and filtering data.

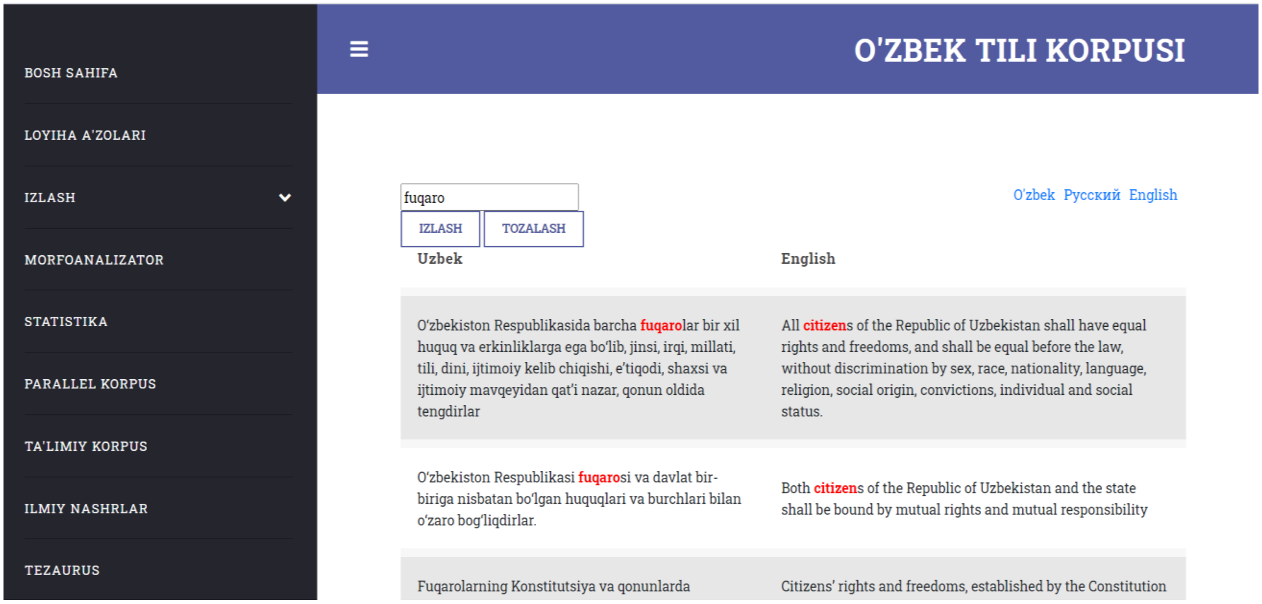
After this step, the following models will be launched: Single Page Model, Search Model, Query Model (the request is sent directly to the Search Model system, not from the controller), Context Model, Single Page Edit Model, Statistics Model, and Data Management Model. Functional models such as Document Model, Sentence Model, Word Model, Security Model were also used effectively [5].

Uzbek language corpus manager is available at https://uzbekcorpus.uz (Pic.1). Formal-functional model of the Uzbek language corpus consists of architecture of the text types, corpus manager system, and language analysis. Searching engine is intended for use as objects of various fields. In the field of pedagogy and computer linguistics, in particular, generated texts can be used effectively through a search query provided in the interface using the corpus manager. Search by lemma, token and concordance organized by n-gram by lexical units:

By token search system algorithm count each word by word in lexicon and rest of parts analyzed at tokenization stage. In Uzbek words might be derivational and grammatical forms:

[gul]=>{gulchilik-gulli-guldor, gulzor}=>derivation words
[gul]=>{gullar; gulag; gulgami; gulniki, gullardanmi…}=>grammatical categorization

Uzbek corpus consists of subcorpora: parallel corpus and learner corpus. Parallel corpus includes Uzbek-English parallel texts in official and scientific style. After segmentation text and word forms alignments input database (pic.2)



Uzbek	English
O‘zbekiston Respublikasida barcha fuqarolar bir xil huquq va erkinliklarga ega bo‘lib, jinsi, irqi, millati, tili, dini, ijtimoiy kelib chiqishi, e’tiqodi, shaxsi va ijtimoiy mavqeyidan qat’i nazar, qonun oldida tengdirlar	All citizens of the Republic of Uzbekistan shall have equal rights and freedoms, and shall be equal before the law, without discrimination by sex, race, nationality, language, religion, social origin, convictions, individual and social status.
O‘zbekiston Respublikasi fuqarosi va davlat bir-biriga nisbatan bo‘lgan huquqlari va burchlari bilan o‘zaro bog‘liqdirlar.	Both citizens of the Republic of Uzbekistan and the state shall be bound by mutual rights and mutual responsibility
Fuqarolarning Konstitutsiya va qonunlarda	Citizens’ rights and freedoms, established by the Constitution

Picture 2. Subcorpus – Parallel corpus between Uzbek and English

The use of a translation memory environment not only speeds up the human translation process, but also has a positive effect on its quality. This is because usage the features of this program could give opportunity to show mistakes and correct them in the text, as well as to match the meanings of different styles of bilingual texts.

The definition of concordances in parallel texts is based on the alternative equivalence of a word, phrase, or stable combination in a given language.

In this case, adequate translation does not always justify itself, as some words are dropped or a component is added to change their lexical and grammatical model. Therefore, for keywords in parallel texts, active words and terms that are frequently encountered in the context and have not undergone specific changes in translation are aggregated into a database.

In this case, auxiliary word groups in Uzbek (connective, auxiliary and preposition), words that do not have an independent meaning (imitation, adverb, auxiliary verbs, independent verbs, etc.) and we will exclude from database the words which are often used as homonyms in the text.

Because it is impossible to predict their meaning in translation. The number of compound words in Uzbek language is relatively large and can be as follows: {n… S / S… n1} => compound word (point of view / disregard) {n… SWn} => phrase (look down)

For example, the following keywords are considered as frequently used normative templates in abstracts: word base => derivation form => word form => frequency of frequently used words, phrases or terms => concordance => translation. The text in the sample is a "long" sentence because it has a large number of punctuation and predicative elements, and the text has been translated into two parts by the translator. The generation of parallel sentences separated from the large text context can be done in the following steps [9]: Information about alternative pairs of tokenized words from parallel text is obtained; A long sentence is divided into separate segment sentences up to the part with certain punctuation marks “, ”, “; ”, “: ”; Basis - the amount of content of the segments of the translated sentences is counted and the state of conformity is determined; Translation - the alternative status of the main texts is determined; If more or more relationships are observed between segments, several segment units are interconnected or attached in the form of a single relationship.

Learner corpus is focused to study lexical minimum of school pupil or language acquisition Uzbek as a second language. It has literatures 5-11 classes with metadata. Corpus is for specialized purposes corpus 3content scientific and publicity style. Here is the statistics of texts:

Style	**Words**	**Lemma**	**token**	**Sentence**
Fiction	524120	25032	251652	42032
Science	321402	15401	245635	21000
Publicity	120232	12354	120123	10000
Official	165239	8652	965861	23007

Our corpus includes language analysis system: morphoanalyzer, parsing, and semantic analysis. Linguistic annotation as a level of morphology we intended to tag universal markup tagging system. Therefore, there are two graphemes of parts of speech and their POS tagging set. Search system of corpus is enclosed search system by lemma, token, KWIK and collocation.

Preparation of language query system for corpus as mentioned before we included 85 thousand words in Uzbek lexicon for both graphemes.

Here is showed POS tagging set for morphology:

Example of POS	**TAG**	**National name**
o'quvchi (pupil)	NOUN	Ot
bor (go)	VERB	Fe'l
go'zal (beatiful)	ADJ.	Sifat
tez (fast)	ADV.	Ravish
bir (one)	NUM.	Son
hamma (all)	P	Olmosh
o'qish (reading)	V_N	Harakat nomi
kulayotgan (laughing)	V_S	Sifatdosh
borgiz (s+make go)	V_O	Orttirma nosbatdagi fe'l
ko'ril (obj+was seen)	V_PASS	Majhul nosbatdagi fe'l
orqali (by, through)	ADP	Ko'makchi
-mi (question meaning of the word)	PART	Yuklama
agar (if)	CONJ	Bog'lovchi
vov (av-av)	Imit.	Taqlid so'z
ehtimol (probably)	MW	Modal so'z
voy! (wow)	EXL	Undov so'z

Concordance is searched by left and right side algorithm. Due to the corpus contained documents in both Cyrillic and Latin alphabets, lexicon was also stored in two alphabets. The algorithm for finding the lemma is shown in the example of the Latin alphabet. a list that contains all the words of the Latin variable. Looking for a lemma searched from left to right at the beginning of a word, the rest of the word is searched within the suffix models corresponding to the current word (suffix models are stored in the list of suffixes). The lemma is considered correct only if the search among the attachment models is successful result.

```
def findLemma(word):
    lemma = word
    for i in range(1, len(word)):
        if (word[:i] + '\n' in lotin):
            for st in range(len(suffixes)):
                if(suffixes[st]==word[i:]):
                    print("suffix :", word[i:])
                    ind=forms[st].find(':')
                    print('form: ', word[:i]+forms[st][6:ind+1]+word)
                    lemma = word[:i]
                    return lemma
    return word
```

For morphological analysis we used HFST technology and we obtained morphotactic structure of the words: In order to morphological analysis there are three components of the Uzbek language: alphabet (Latin and Cyrillic), grammatical rules and Lexicon. In Uzbek the following morphotactics of words as example of Noun:

```
LEXICON NumC
+SG: Poss1;
 +PL: lar Poss2;
LEXICON NumV
+SG: Poss2;
+PL: lar Poss2;
LEXICON
Poss1 +PP1+PSG: m Case;
+PP2+PSG: ng Case;
+PP3+PSG: si Case;
+PP1+PPL: miz Case;
 +PP2+PPL: ngiz Case;
 +PP3+PPL: i Case;
0:0 Case;

LEXICON

 CyrPrePrefinal1 0:0 Final;
+PART: ми Final;
+PART: ку Final;
+PART:-^Ya Final;
+PART:-да Final;
+PART:-чи Final;
```

V. **CONCLUSION**

Corpus morphoanalyzer and parsing as a tool is as an object of investigation for NLP or computational linguistics will help to enhance researches of computational linguistics in perspective. Additionally, it is also play crucial role to investigate other field of linguistics and humanities and social sciences. As the first corpus of the Uzbek language has perspective and a number of investigations will be conducted in this sphere in the future.

REFERENCES

[1] Sulevmanov, D., Gatiatullin, A., Prokopyev, N., Abdurakhmonova, N. (2020) Turkic morpheme web portal as a platform for turkology research International Conference on Information Science and Communications Technologies, ICISCT 2020, 2020, 9351500.

[2] Khusainov, A., Suleymanov, D., Gilmullin, R., Minsafina, A., Kubedinova, L., Abdurakhmonova. N. (2020) First Results of the "TurkLang-7" Project: Creating Russian-Turkic Parallel Corpora and MT Systems CMLS 2020 CEUR Workshop Proceedings, 2020, pp. 90-101.

[3] Khusainov, A., Suleymanov, D., Gilmullin, R., Gatiatullin, A. (2018) Building the Tatar-Russian NMT system based on re-translation of multilingual data Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11107 LNAI, pp. 163–170.

[4] Абдураҳмонова Н. (2020) Замонавий корпусларнинг компьютер моделлари // Ўзбекистонда хорижий тиллар. -2020. - № 1(30). - Б. 50-58. https://doi.org/ 10.36078/

[5] Мухамедшин, Д.Р., Сулейманов Д.Ш. (2018) Система корпус-менеджер: архитектура и модели корпусных данных Программные продукты и системы / Software & Systems 4 (31) – С. 6.

[6] В. П. Захаров, И. В. Азарова, О. А. Митрофанова, А. М. Попов, М. В. Хохлова (2019) Моделирование в корпусной лингвистике Специализированные корпусы русского языка, Санкт-Петербургский государственный университет. -С. 19.

[7] Erhard Hinrichs, Marie Hinrichs, Thomas Zastrow, Gerhard Heyer, Volker Boehlke, Uwe Quasthoff, Helmut Schmid, Ulrich Heid, Fabienne Fritzinger, Alexander Siebert, and Jorg Didakowski. (2009) Weblicht: Web-based LRT services for German. In Workshop on linguistic processing pipelines, GSCL Jahrestagung, Potsdam.

[8] Аброскин А. А. Поиск по корпусу: проблемы и методы их решения // Национальный корпус русского языка: 2006-2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009, 277-282.

[9] https://uz.wikipedia.org/wiki/O%CA%BBzbek_tili

[10] Jinyi Zhang, Tadahiro Matsumoto (2019) Corpus Augmentation for Neural Machine Translation with ChineseJapanese Parallel Corpora / Applied sciences (9), 2036.