BASICS OF COMPOSITION OF THE CORPUS OF LACUNAR UNITS IN UZBEK DIALECTS

Abdurahmanova Sayyora Bozorboy kizi
sayyoraabdurahman@gmail.com
PhD TSUULL

ABSTRACT

The article discusses the basics of including lacunar units in Uzbek dialects into the dialectal corpus. At the same time, the importance of lacunar units and the issues of enrichment due to internal resources were specially addressed.

Key words: *corpus, dialect, lacuna, internal source, external source, historical source, vocabulary, dialectal corpus, dialect, language, lemma, token, lemmatization.*

АННОТАЦИЯ

В статье рассматриваются основы включения лакунарных единиц в узбекских диалектах в диалектный корпус. При этом специально затрагивались важность лакунарных единиц и вопросы обогащения за счет внутренних ресурсов.

Ключевые слова: *корпус, диалект, лакуна, внутренний источник, внешний источник, исторический источник, лексика, диалектный корпус, диалект, язык, лемма, лексема, лемматизация.*

INTRODUCTION

A corpus is a set of texts submitted to a search program in order to determine the characteristics of language units, a set of written, spoken, and computerized search texts stored in a language in electronic form. From the point of view of linguistics or language, the body of the corpus of texts is understood as a large, machine-readable, single, systematic, marked, philologically qualified language complex and a database providing specific information on linguistics. E. Finegan defines: "A corpus is a collection of texts, usually in a computer-readable format, that gives us information about the context, informant, author, addressee, or audience in which the text was produced." says. Various general informational social sites are a large, systematic collection of texts that can justify the language rules and phenomena encountered in the identified fields on the basis of statistical analysis and hypothesis testing [Zakharov V.P. 2011:7] describes as.

Based on this, it should be said that language corpora are necessary, first of all, for a linguist. After all, it is difficult to imagine today's theoretical and practical philology without a corpus.

Oriental Renaissance: Innovative,
educational, natural and social sciences

SJIF 2023 = 6.131 / ASI Factor = 1.7

(E)ISSN:2181-1784
www.oriens.uz
3(6), June, 2023

In cases where evidence is used in linguistic research, those facts should be collected and systematized. We need a national corpus to improve the effectiveness of such research. The dialectal corpus, which is part of the national corpus, reveals the diversity of the language. We focused our attention on lacunar units in Uzbek dialects. The main reason for this is that lacunar units are the richness of our language. To make it more clear, it is the lacunar units that serve to ensure the accuracy of word acquisition, phonetic phenomena, historicity, semantics, and inclusion in the literary language, as well as statistical research.

DISCUSSION AND RESULTS

Corpus can be created for different purposes and in different fields. According to the specialization of the text, dialectal corpora are a separate group, which, in turn, can include dialects, phonetic, lexical features of dialects, and lacunar units in dialects.

The corpus of dialectal texts - it contains records of dialectal speech of different regions of Uzbekistan with their grammatical specification preserved, a special search taking into account dialectal morphology is taken into account. In turn, corpora are divided into such types as monolingual, bilingual or multilingual, oral and written. In a monolingual corpus, language variants and dialects are contrasted with each other, while a bilingual and multilingual corpus consists of a collection of texts written in different languages within the same topic. Multilingual corpora are mainly used by translators. By creating a dialectal corpus, we can collect dialectal materials, and by placing lacunar units in the dialectal corpus, we can introduce the alternative of concepts and words that are not available in the language into the literary language.

In turn, it is important to group these lacunar units before including them in the corpus. After that, when creating a corpus, we must first determine some principles. First of all, materials for the hull are collected, then they are sorted and processed. After completing these processes, we pass each lacunar unit to the stage of lemmatization and tagging. With the participation of these units, we will also have to put sentences and phrases into the program.

This process is not difficult because the phenomenon of homonymy is almost never found in lacunar units. In creating the corpus of lacunae, we use the dictionary of lacunae in the Uzbek language, dictionaries of Uzbek dialects, and sources collected in dialectological practices.

Corpus data is distributed on the basis of a license in accordance with the laws of the Republic of Uzbekistan. They also indicate the source of the materials.

Oriental Renaissance: Innovative,
educational, natural and social sciences
SJIF 2023 = 6.131   / ASI Factor = 1.7

(E)ISSN:2181-1784
www.oriens.uz
3(6), June, 2023

so'z	izohi	turkum					
		ot	sifat	son	fe'l	olmosh	ravish
ĀBDŎL TĪLI (абдол тили)	xalq san'atkorlarining yashirin tili. Bu "til" yuzlab so'z va so'z birikmalarini qamrab oladi. Masalan, paypoq maxsum – tomosha paytida tushgan puldan yashirib qoluvchi shaxs, galstuk – o'qituvchi (Xorazm, Gurlan shevasi; S.A.)						
ĀBIRŽI (абиржи	bahor paytida daryodagi suvning muz aralash oqimi (Janubiy Qoraqalpog'iston o'zbek shevasi; O'.E).						
ĀBŽEŠ (абжеш	– maska uzumdan tayyorlangan ichi danakli mayiz (Janubiy Qoraqalpog'iston o'zbek shevasi; O'.E).						
ĀBŽOŠI (абжўши	ishqor solingan qaynoq suvga botirib, so'ng quritilgan mayiz (Qashqadaryo,Maymanaq;N.T).						
ĀDRĀMKAŠ (адрамкаш)	uchi ilmoqli katta juvoldiz (Xorazm; A.A) .						
ĀDIŠĀK (адишак)	yovuz maxluq ko'rinishidagi afsonaviy kuch (Qashqadaryo, Chiroqchi; N.T).						
ĀFŎN - SAFŎN (афон-сафон	– biron katta ishning yakunlanganligi, tugaganligi (Turk - kaltatoy shevasidagi ayrim so'zlar. Baxmal						
ĀJYĪZ (айғиз	aravaning yuk ortiladigan qismi (O'zbek xalq shevalari lug'ati. Andijon, 1971; To'plovchi: H.A.).						
ĀJMŎQĪ (аймоқи)	– ko'milmaydigan tok (O'zbek xalq shevalari lug'ati. Toshkent, 1971; To'plovchi: H.A.).						
ĀJIQQULOQ (айиққулоқ)	oshqozon kasalligini davolashda ishlatiladigan, karamga o'xshash bargli dorivor tog' o'simligi						
ĀŽIVĀQA (аживақа)	O'smay, rivojlanmay qolgan; majmag'il. G'anakni almashtirish vaqtida mana bundaqa ajivaqalarni – o'smay qolgan qurtlarni olib ko'rsatdi. Oliyaxon. N. Safarov.						
ĀJNIMŎL (айнимол)	fe'l - atvorida turg'unlik xislatlari bo'lmagan (odam) (Qashqadaryo, Uyshun; N.T).						
ĀJRĀM (айрам)	oqar suvning chuqur yeri (Buxoro, Qorako'l; A.A).						
ĀJRIĆOKI (айричўки)	tikanli xashak (yantoq, olako'ztikan, jinjak)larni o'rishda ishlatiladigan uchi ayri cho'p (Qashqadaryo,						
ĀJRILMĀ (айрилма)	katta ariqdan chiqarilgan , ayrilgan (Qashqadaryo, Qorashina; N.T).						
ĀJRITĀJOQ (айритаёк	– ikki tayoqning bir xil uchini birlashtirib mixlab, sim bilan o'rab qotirilgan, quyi qismi orasi ikki metrdan iborat yer o'lchov moslamasi (Qashqadaryo, Langar; N.T).						
ĀJRITŮJOQ (айритуёк)	1) tuyoq orasining kasallanishi; 2) ikki tuyoqli (mol: qo'y, echki) (Qashqadaryo, Ko'kabuloq; N.T).						
ĀJTĪÝLĪ (айтигли	– aytishga arziydigan, aytsa bo'ladigan (Qashqadaryo, Zachkana; N.T).						
ĀJYĪZ (айғиз) –	aravaning (novdadan to'qilgan) yuk ortadigan qismi (Andijon; S.N).						
ĀŽI (ажи)	o'ynaladigan oshiqning yutuq deb hisoblanadigan tomoni (Qashqadaryo, Dehqonobod; N.T).						
ĀŽIDI (ажиди)	belgilangan chiziq ustiga ost-ust qilib terilgan oshiqni soqqa bilan ma'lum masofadan urib yiqitish va marradan chiqarish, o'yin (Qashqadaryo, Toshqo'rg'on; N.T).						

a. classification of words to be placed in the corpus of lacunae.

so'z	izohi	qo'shimchalar								
		-ning	-ni	-ga	-da	-dan	-m	-im	-ng	-i
ĀBDŎL TĪLI (абдол тили)	xalq san'atkorlarining yashirin tili. Bu "til" yuzlab so'z va so'z birikmalarini qamrab oladi. Masalan, paypoq maxsum – tomosha paytida tushgan puldan yashirib qoluvchi shaxs, galstuk – o'qituvchi (Xorazm, Gurlan shevasi; S.A).									
ĀBIRŽI (абиржи	bahor paytida daryodagi suvning muz aralash oqimi (Janubiy Qoraqalpog'iston o'zbek shevasi; O'.E).									
ĀBŽEŠ (абжеш	– maska uzumdan tayyorlangan ichi danakli mayiz (Janubiy Qoraqalpog'iston o'zbek shevasi; O'.E).									
ĀBŽOŠI (абжўши	ishqor solingan qaynoq suvga botirib, so'ng quritilgan mayiz (Qashqadaryo,Maymanaq;N.T).									
ĀDRĀMKAŠ (адрамкаш)	uchi ilmoqli katta juvoldiz (Xorazm; A.A) .									
ĀDIŠĀK (адишак)	yovuz maxluq ko'rinishidagi afsonaviy kuch (Qashqadaryo, Chiroqchi; N.T).									
ĀFŎN - SAFŎN (афон-сафон	– biron katta ishning yakunlanganligi, tugaganligi (Turk - kaltatoy shevasidagi ayrim so'zlar. Baxmal tumani; H.A.).									
ĀJYĪZ (айғиз	aravaning yuk ortiladigan qismi (O'zbek xalq shevalari lug'ati. Andijon, 1971; To'plovchi: H.A.).									
ĀJMŎQĪ (аймоқи)	– ko'milmaydigan tok (O'zbek xalq shevalari lug'ati. Toshkent, 1971; To'plovchi: H.A.).									
ĀJIQQULOQ (айиққулоқ)	oshqozon kasalligini davolashda ishlatiladigan, karamga o'xshash bargli dorivor tog' o'simligi (Qashqadaryo, Dehqonobod; N.T).									
ĀŽIVĀQA (аживақа)	O'smay, rivojlanmay qolgan; majmag'il. G'anakni almashtirish vaqtida mana bundaqa ajivaqalarni – o'smay qolgan qurtlarni olib ko'rsatdi. Oliyaxon. N. Safarov.									

b. separating the words to be placed in the corpus of lacunas according to the acceptance of suffixes.

Creating a corpus of lacunar units in the Uzbek language is very important:

- Uzbek language vocabulary is collected;
- serves to enrich the language due to its internal capabilities.
- helps to make historical sources understandable.
- reveals the subtlety of meaning.
- serves to analyze the real speech situation.

The corpora are divided into simple and descriptive language corpora according to their characteristics and the basic level for research. So, we need to clarify what the described corpus will be. Linguistically described or marked can be interpreted as a non-textual, non-textual, but somehow informative resource for the corpus. We use this to facilitate the work of the automatic analysis corpus. When first marking the corpus of a language, we can insert any symbol, as long as there is some relation between the symbol and the word. Then we won't have any difficulty in finding words similar to this word. Markup helps you to identify a specific place in the text.

REFERENCES

1. Xolmanova Z. O'zbek shevalaridagi lakunlar lug'ati. Toshkent 2022

2. Abdullayeva O. O'zbek tilidagi internet matnlari korpusini yaratish f.f.f.d., PhD dissertatsiyasi Toshkent 2021

3. Z. Xolmanova, O. Saidahmedova, O. Nurullayeva Lingvokulturologiyaga oid tushunchalar tadqiqi – Toshkent 2018

4. Кутузов А.Б. Корпусная лингвистика. – М., 2005. С. 15-16.

5. Захаров В.П. Корпусная лингвистика. – Иркутск, 2011.С. 7.