

THEMATIC AND PASSPORT META-MARKUPS OF DIALECTAL TEXTS IN UZBEK NATIONAL CORPUS

(on the example of saying “j” dialects in Boysun district)

Kholova Muyassar Abdulhakimovna

Teacher of the Department of Uzbek Linguistics, Termez State University



Crossref <http://dx.doi.org/10.37057/2433-202x>

Issue DOI <http://dx.doi.org/10.37057/2433-202x-209-2020-1-6>



Article DOI <http://dx.doi.org/10.37057/2433-202x-2020-1-6-26>

Abstract: This article discusses the issues of the language corpuses of the world, the role of dialect texts in the Uzbek National Corpus, the creation of a correct, easy and convenient search engine for the necessary information, thematic and passport meta-markuping of large texts.

Keywords: language corpus, Uzbek national corpus, dialectal text, thematic meta-markup, passport meta-markup

Introduction

The first linguistic corpus began to appear in the 60s of the last century, in 1963 at Brown University, USA, when the first large corpus (Brown Corpus) was created in science, the dialect corpus, which is considered as our research object, can be divided into 2:

1. Dialect corpuses created on the basis of formal features of other language corpuses, taking into account the peculiarities of their dialects: Portuguese corpus - The syntax-oriented Corpus of Portuguese dialects [14], Polish corpus – Dialecty I gwary polskie Kompendium internetowe [15], Spanish corpus–Corpus Oral y Sonoro del Español Rural[16], Bulgarian corpus – Bulgarian dialectology as Living Tradition[17], Georgia corpus (consisting of 2 054 705 words,

3400 texts)[18] , Mandarin corpus by the Chinese language (this dialect reflects the diversity and variability in the Chinese language) and so on.

2. Dialect corpus based on the dialect of a certain region: Kuban dialect corpus [4, 67], Saratov dialect corpus [10], Dialect corpus of Ustyia district of Arkhangelsk region [11],

The corpus of dialect texts of the National Corpus of the Russian Language [13] is still among the corpus in which free use is supposed. It would be more correct to divide the studies on this corpus into two stages.

The first stage began in 2006, and at the first stage grammatical meta-marked spelling standards were given in the abdominal cavity, but from such texts there is always a complete database.

At the second stage, researchers working on the corpus developed new, completely new for their time, standards for the presentation of dialect texts. The updated dialect corpus includes 285 thousand words collected from 22 regions of Russia, 627 texts, of which 15 text audio recordings, 126 non-spelling texts, 431 meta-markups with phonetic markings indicating the person providing detailed information (this information was obtained as of 2016. 9. 12) [13]. Initially, a site [<http://mitruis.ru/dialect.html>] was developed for dialectology researchers with the goal of using and downloading metatexts and grammatical markup texts [3, 149]. Of the dialect texts, attention was paid to the writing of transcriptions, emphasis on words and suffixes.

Russian researchers recommend conducting dialect text markup in three directions:

- a) search-address (in this case, the questionnaire for any text is issued in the following sizes);
- B) according to the phonetic level (in the phonetic analysis of dialect texts, the user puts a mark indicating that in spelling form to distract attention);
- c) thematic-genre.

The correct determination of the direction of the corpus of Uzbek dialects is one of the main parameters for conducting dialectological studies. Some stages of the development of linguistics can be traced in copyright [8] and synonyms [1], which are created to this day. In the dialectological corpus, through the anthropocentric questionnaire, the data warehouse is gradually expanding, thanks to which it becomes possible to use full texts. The Bulgarian, Scandinavian, Estonian, Portuguese, Scottish and Tomsk dialectological corpuses, along with the texts contained in the electronic library, have the opportunity to use a corpus consisting of a set of dialects in a centralized form [5]. And in the Russian National Corpus (RNC), it would initially be possible to work with

texts, delivering fragments of texts, and then through a direct survey directly [3,162]. Creating an electronic format of dialectical materials, scanning them, for a wide and free audience

To replenish the data warehouse of the dialectological corpus, work is underway on dialectological dictionaries (52,000 dialect units, including the Bashkir dialect) [7]. The dictionary also contains phonetic, semantic and morphological dialectisms with words inherent in a special syllable. The use and search of the database is carried out in the following steps:

1. a word related to dialect
2. a phrase
3. dialect
4. sheva (dialect)
5. literary norm
6. translation into Russian

For example: имән бармак – исем, мөньяк-көнбайыш диалекты, танып һойләше, һик бармак, указательный палец [7]; Тажар [a word in a dialect] – noun, kipchak dialect [dialekt], Boysun dialect (Kolkamish, Khojabulgon) [dialect], хачир [literary norm], осёл [translation into Russian].

The search system in the database is performed in the same six fields, and the query is given only in the form of a database of lexemes. It is advisable to conduct the survey as follows:

- by entering a clear dialectal word;
- by entering the dialectal word that comes at the beginning of the sentence;
- by entering the dialectal word that comes at the end of the sentence;
- by entering the corresponding word in any part of the sentence.

Questioning a word in a variety of ways allows for a wider use of the underlying source and serves for the accuracy of the researchers' conclusions. In the Bashkir dialect there are 148 dialect variants of khankizi (ladybird), 17 of which are found only in the eastern dialects [7].

The passport part of the meta-markups (dialect text) contains the text as follow.

| t/r | Survey | Answer |
|-----|---|--|
| | Recorded location (complete information makes searching easier) | District, region, republic |
| | Recorded time | year, day, month |
| | Informant | Last name, first name, patronymic name |
| | Gender of the informant | Gender |
| | Place of birth of the informant | District, region, republic |

| | | |
|--|-------------------------|---|
| | Degree of the informant | Uneducated, higher educated (bachelor, master, PhD) |
| | Data collector | Last name, first name, patronymic name Institution where he/she works (for example: TerSU) |
| | Notebook or information | № ____ |
| | Topic | Family, holiday, work |
| | Reality period | Until independence; After independence |

The description of events refers to a period of time (30s, 50s, 60s, etc.) that indicates the possession of a specific thematic characteristic. The information is presented in a single concept that unites the perceptions of the world, the themes. An analysis of the records shows that the dialectal units written in the language of the older generation reflect the traditional culture of the peasants. The connection between the daily life of the ordinary peasant people, their life experiences, and harmony is the basis for the nominative nature of the concepts. Here is a complete thematic list of the Tomsk dialectal corpus: Mehnat qurollari (Working tools). Kasb-hunar (Profession). Hunarmandchilik (Craft). Bog' (Garden). Qishloq tarixi (History of the village). Ish (Work). Uy qurilishi (Construction of a home). Kolxoz (Collective farm). Oziqa tayyorlash (Food preparation). Zig'ir (Flax). Kartoshka (Potatoes). Chorvachilik (Livestock). Asalarichilik (Beekeeping). Kedr ishlab chiqarish sanoati (Cedar manufacturing industry). Ov (Hunting). Baliq ovi (Fishing). O'rnon (o'tin ishlab chiqarish) sanoati (wood production). Qurilish (Construction). Qulupnay (Strawberry). Qo'ziqorin (Mushroom). O'simlik va gullar (Plants and flowers). Taom (Food). Kasallik va da'volanish (Illness and treatment). Kiyim (Clothes). Idish-tovoq (Dishes). Sarxushlik (Intoxication). Sehr-jodu va e'tiqod (Magic and faith). Ibodatxona (Temple). Bashoratlar (Predictions). Qudachilik. Marosimlar (Rituals). To'y (Wedding). Aza (Mourning). Cho'qintish (Baptism). Bayramlar (Holidays). Ob-havo (Weather). Ko'ngil ochish (Entertainment). Siyosat (Politics). Tabiat (Nature). Hodisalar (Events). Zamonaviylik bo'yog'i. Insonni qadrlash (Respect for the person) [6].

Based on the thematic division of the above linguistic corpus, all the topics listed in them cannot reflect the lifestyle of the national ethnic people (Kedr ishlab chiqarish sanoati. Baliq ovi. Qulupnay. Qo'ziqorin. Sarxushlik. Ibodatxona. Cho'qintiish. Zamonaviylik bo'yog'i), the following thematic interface refers to the shelves:

| Social vocabulary | | | Economic vocabulary | | Political- education al vocabulary | | | | | |
|---|---------------------------|----------------------------|-------------------------|---|---|---|------------------|----------|----------------------------------|--------------|
| The concept related to a man and his life | Things and their concepts | | The concept of rituals | National economy | Agriculture | 1. The concept of government | | | | |
| 1. Body parts | Living things | Inanimate things | 1. Game ceremonies | Crafts | Livestock | 2. concept of farming | | | | |
| 2. Concept related to a family and relationship | 1. Camels | 1. Household items | 2. Wedding ceremonies | a. kulolchilik b.mahsidozlik c. coppersmith d. blacksmithing | 2. farming | concept related to the education | | | | |
| | 2. Horses | 2. equipment of otov | | | | | | | | |
| 3. Concept related to illness | 3. Cattle | 3. food and drinks | 3. Mourning ceremonies | 2. needlework | 3. gardening | 4. concept related to the sport | | | | |
| 4. Spiritual experiences | 4. Small ungulates | 4. Building and apartments | 4. Religious ceremonies | a. kashtachilik b. doppidozlik c. chophonchilik d. gilamdozlik e chirozdozlik | | Concept related to the field of justice | | | | |
| | | | | | | | a. Goat | b. Sheep | 5.concept related to the clothes | 4. Farming |
| | | | | | | | 5 .Other animals | | 6. plants | 5. Livestock |
| | | | | | | | 6. Birds | | | |

The conclusion to be drawn from the above considerations is that the expected results for the dialectal corpus are as follows:

- integrating the collected data into a system that ensures the speed of use;
- thematic meta-assessment of texts ensures the accuracy of conclusions of researchers related to traditional and everyday life;
- passport meta-markup of thematic texts provide a complete report on the data collector and informant (informant and informant) and serve as a basis for inductive, deductive conclusions related to the regional dialect.

References

1. A. Eshmominov. Base of synonyms of the Uzbek National Corpus. Dis.- Karshi. 2018
2. I.V. Kachinskaya., D.V. Sichinava. On the corpus of dialect texts in the national corpus of the Russian language. Journals.tsu.ru>lex. 2017.
3. I. B. Kachinskaya., D. V. Sichinava. Dialect subcorpus today // T. A-this is Russian to them. V.V. Vinogradova RAS. 2015.No 6 (6). - S. 142-163.
4. Tregubova E. N. Multi-level thematic markup as a tool for ethnolinguistic representation of dialect discourse in an e-text corpus \\ Vestnik. Tomsk State. University. Phil. No. 1 (33). 2015 pp. 66-77.
5. Yu. V. Ivanov. Tomsk dialect corpus: substantiation of the concept with a development perspective. lexical project and work. 2017 year.
6. E. A. Yurina. Tomsk dialect corpus: at the beginning of the journey. Bulletin of Tomsk State University. No. 2 (14). 2011.
7. Zinnur A. Sirazitdinov, Lily A. Buskunbayeva, Ramiya N. Karimova. Dialectological Base of the Machine Fund of the Bashkir Language as a Tool Research Bashkir Dialect Continuum. Journal homepage: <http://kirigan.com/pubs/vestnik> Vol.23, Is 1, pp. 212-219, 2016. DoI 10.22162/2075-7794-2016-23-1-212-219.
8. Sh. Hamroyava. Linguistic bases of creation of the Uzbek language author's corpus. Dis. - Bukhara. 2018
9. <http://www.dictionary.cambridge.org/dictionary/english-chiese-simplified/corpus>
11. http://sarteorlingv.narod.ru/dialect/elektr_korpus/html
12. <http://parasolcorpus.org/pushkino/index/php>
13. www.ruscopora.ru/search-dialect.html.
14. [http:// www.ruscopora.ru/new/](http://www.ruscopora.ru/new/)