Journal of Advanced Zoology

ISSN: 0253-7214
*Volume **44** Issue **03 Year 2023** Page **1340:1345***

Corpus Linguistics: Study of Folk Parems with The Participation of Zoonyms

Guli Toirova[1*], Guzal Malikova[2], Zarina Komilova[3], Gulasal Khayrulloyeva[4]

[1,2,3,4]Bukhara State University, Bukhara City, M. Iqbol, 11, 200100, Uzbekistan

**Corresponding author's: Guli Toirova*

Article History

Received: 06 June 2023
Revised: 05 Sept 2023
Accepted: 17 Nov 2023

Abstract

The term corpus linguistics is now very popular. The compilation of a corpus of language texts is included among the priority areas of work of the academies. In paremes formed on the basis of the name "insect," homonymy is observed between words that resemble the name of the insect and units that are homophones with the lexeme denoting the name of the insect. Direct homonymy is expressed by the lexemes mole, worm, donkey (dial. scorpion), and "burga" - "burgan" act as homophones. The creation of text corpora is considered by a number of scientists as the most important humanitarian task of linguistics. This article explains the concept of corpus linguistics and discusses its theoretical foundations.

Keywords: *Corpus Linguistics, Linguistics, Method, Teaching, Computer Technologies, Language, Insect, Direct Homonymy Worm, Donkey (Dial. Scorpion).*

1. Introduction

The field of corpus linguistics includes all linguistic research based on the material of a corpus of texts. We will try to give a definition of a corpus a little later, but for now we note that corpus linguistics is not a direction associated with a certain tier of the language system (like phonetics, lexicology or syntax), or a certain theory (like functional or generative grammar), or aspect of analysis (formal, semantic or pragmatic). It is rather an ideology according to which the results of linguistic research should be based primarily on the analysis of texts (oral or written), and not on the intuition of the researcher or informant.

Apparently, there are not many supporters of the radical approach who completely deny the role of intuition. For linguists who consider themselves to be part of the corpus direction, we are talking specifically about a system of priorities: any conclusion must be confirmed by the material of "natural" texts, and not just by judgments about the acceptability of a particular construction obtained in the conditions of a linguistic experiment [1].

The article by N.V. Zimovets on the linguistic formation of proverbs [10], the monograph by I. Sirota [11], the dissertations of B.B. Mansurov and S. Basharan [12] consider the factors of the emergence and formation of proverbs, and this direction is being studied today, the most important issues. In recent years, in Uzbek linguistics, doctoral dissertations have been defended by B. Juraeva on the topic "Linguistic foundations of the formation of Uzbek folk proverbs", D. Turdalieva "Linguopaetic features of Uzbek folk proverbs", D. Tosheva "Linguocultural characteristics of proverbs with a zoonymic component", "Euphemization of folk proverbs in the Uzbek linguistic and cultural environment" Sh. Kalandarova[13, 14].

2. Materials And Methods

Corpus linguistics is the study and analysis of data obtained from a corpus. The main task of the corpus linguist is not to find the data but to analyse it. Computers are useful, and sometimes indispensable, tools used in this process.

Corpus-based studies involve the investigation of corpora, i.e. collections of (pieces of) texts that have been gathered according to specific criteria and are generally analysed automatically. Defining and Developing Translation Competence for Didactic Purposes: Some Insights from Product-Oriented Research [2].

A corpus can help us identify terms shown in context, and the most frequent patterns of use. From the different concordance lines, collocates and clusters (retrieved thanks to the software Concord, a functionality provided by WordSmith Tools), we obtain relevant grammatical and lexicographical information.

Corpora have not only been used for linguistics research, they have also been used to compile dictionaries (starting with The American Heritage Dictionary of the English Language in 1969) and grammar guides, such as A Comprehensive Grammar of the English Language, published in 1985.

Entomology is a science that studies insect species and their beneficial and harmful aspects, methods and technologies for their practical use. Entomology (Latin entos - insect and logia - science) is the science of insects, studying the structure, life of insects, their individual and historical development, diversity, distribution on earth, their connection with the environment, etc. According to the task, they distinguish between theoretical, that is, general entomology and applied entomology. General entomology is divided into insect morphology, embryology, physiology, biochemistry, ethology, entomogeography, paleontology, systematics and other sciences. These subjects can be divided into smaller sections depending on the subject of study. For example, in taxonomy, coleopterology studies coleopterous birds, lepidopterology studies butterflies, and myrmicology studies ants.

What is a corpus?! In a certain sense, the overwhelming majority of modern linguistic research (with the exception of purely abstract theories such as glossematics or early generativism) is based in one way or another on textual material. Probably all linguists had to work with cards or with electronic records (transcriptions) of texts. If the conclusions of a study are based entirely on clearly defined textual material, this material can be called a corpus. The only question is how indicative (representative) this corpus is for judging the language as a whole.

It is common to distinguish between a corpus and a collection (or library) of texts. The characteristic features of the corpus are often cited as its large size (tens of millions of word usages) and the presence of linguistic markup.

3. Results and Discussion

In our opinion, the distinctive feature of a corpus is, first of all, its representativeness. At the same time, the size of the corpus that meets the requirement of representativeness depends on the research for which it is intended. For research in the field of phonetics, prosody, morphological typology ("Greenberg indices"), determination of the dominant word order and the most frequent syntactic models, etc. there is no need to involve huge arrays of texts. Representativeness here will be determined by the representation of various functional styles, dialects and sociolects, and a diachronic perspective. However, the corpus of a particular study may well be limited to the framework of one regional or social dialect or even the speech production of an individual.

At the same time, if we are interested in some peripheral phenomena of vocabulary or grammar, the processes of grammaticalization of individual lexemes, the emergence and development of certain syntactic structures, it is necessary to involve a much wider material.

Reference corpus. Ideally, one should strive to create a Corpus of Language (with a capital L) - a corpus that is "representative in all respects", which could serve as a reliable source of data for any linguistic research [3].

In English-language literature, such a corpus is designated by the term reference corpus 'exemplary (?) corpus'. The English scientist J. Sinclair, the author of a programmatic article on the typology of corpora, gives the following definition: "A model corpus is created in order to provide complete information about the language. It must be large enough to represent all the significant varieties of that language and its characteristic layers of vocabulary and thus serve as the basis for grammars, dictionaries and other reliable reference literature."

Of course, a corpus that fully meets the requirement of representativeness is an ideal that is hardly possible to achieve. However, even a distant approach to it gives linguists (and not only linguists!) a powerful tool for studying language (and through language, the culture of a people).

The humanitarian role of the corps. The general humanitarian role of large text corpora seems to be very significant. We can say that a corpus is a new, unique form of language life. Unlike paper files, which, after the completion of the research or publication for which they were intended, at best end up in storage in the archive, the electronic corpus continues to live, be enriched, merge with other corpora and actively serve subsequent generations of computers. Of course, provided that this housing

is designed in such a way that it can be integrated with other housings, and the next revolution in technology does not make it unsuitable for further use.

In addition to solving scientific problems themselves, the corpus of texts can be used for didactic and even purely practical purposes. Anyone who has had to write texts in a non-native language knows the problem: even the best dictionaries with a large number of examples do not always allow one to conclude how "natural" a particular construction sounds and how accurately it reflects the meaning put into it. It's good if you have a native speaker at hand (who also has a good sense of style). Now imagine that we have the opportunity to check whether such a construction occurs in a corpus of texts, and if so, in what context and in what works. Unfortunately, it is not yet possible to realize such a dream in practice. Technically, this would not be difficult, but the existing large corpora of texts are currently closed to free access.

In particular, the lexeme ant as a factor representing a positive characteristic:

1) based on the seme "hard work": Chumoli yuk tashir, Yomon odam gap tashir.

2) based on the seme of "diligence": Tirishqoqlikni chumolidan o'rgan, Dangasalikni qurbaqadan.

3) based on the seme of "friendliness": Chumoli birlashsa, chayonni yiqar. Chumoli birlashsa, chayon po'stini yirtar. Chumoli biriksa, sherni yiqitar. Yetti chumoli birlashib, bir yovni yiqitar.

4) based on the seme of "material security": Chumolining iniga qurbaqa chivin so'rab kelibdi. Chumolidan qurvaqa xayr so'rabdi.

The study revealed 7 types of positive meaning: 1 seme in the lexeme bee, 3 semes in the lexeme ant, 1 seme each in the lexemes louse, butterfly, spider, beetle and fly.

Tools for working with the corpus. In addition to the texts themselves, a full-fledged corpus must have a set of "tools" for working with them. These tools can be divided into two categories [4]:

1) tools for viewing texts and requesting data;

2) means of enriching the corpus with analytical information, which is called annotation, or markup, tagging.

The most common ways of viewing a text are imitation of an edition (with possible selection of objects of interest to the researcher) and concordances (a list of word forms or phrases in context). The main advantage of the electronic publication over the printed one is the ability to quickly search for forms and combinations of interest to the researcher. The breadth of search parameters depends on what analytical information is encoded in the corpus. If we want to find all occurrences of a certain word form, then this can be easily done in a simple text file. If we want to find all occurrences of a certain lexeme represented by a number of word forms, this is somewhat more difficult, but also possible. If we want to find all cases of use of a certain gramme (for example, the instrumental case of the singular of a noun), doing this on an unlabeled corpus is extremely problematic.

What is corpus marking?

As already noted, marking is the enrichment of a corpus with various kinds of analytical information.

The minimum marking, which, as a rule, is easily carried out automatically, consists of equipping the corpus with reference information. In other words, when we receive a response from the corpus to our request, we must clearly know the "coordinates" of our example ("text / chapter / paragraph" or "page / line"). For linguistic research, morphological marking is of great value: each word form is correlated with the "initial" ("dictionary") form of the lexeme, its part-speech affiliation and grammes of inflectional categories are determined.

Automatic morphological marking programs have been developed for many languages, but all of them give one or another percentage of defects (inevitable due to linguistic homonymy) and require "manual" checking. In some cases, however, you can limit yourself to rough automatic data and take into account the percentage of error.

Other types of linguistic markup are also possible: syntactic, semantic, pragmatic, etc., but their universality is not so obvious. If we can talk about a relative consensus among linguists on the issue of the main parts of speech and the composition of grammes, then syntactic functions and semantic groupings are not understood in the same way by different linguistic schools.

Thus, a separate methodological problem arises: how to ensure that differences and even contradictions in linguistic theories and research interests do not interfere with the successful

functioning of the corpus for the benefit of the entire linguistic community? It should immediately be noted that there are technologies that can solve this problem.

In addition to linguistic markings, there is also **philological** marking. It allows you to include text variants, author's and editorial edits in the corpus, highlight foreign words, quotes, direct speech of characters in a literary work, and various kinds of stylistic figures.

Analytical marking of a corpus is a very labor-intensive process, but it is not without scientific interest in itself. In the process of "pasting labels" on word forms or syntactic structures, the "bottlenecks" of the classifications used are revealed, and interesting examples attract attention. And the main thing is that the results of this painstaking work will not gather dust in the archives, but will be actively used and developed.

Insects include invertebrates, which include moths, ants, flies, fleas, lice, spiders, leeches, scorpions, beetles, butterflies, bees, mites, worms, grasshoppers, moths, including mites. They occupy first place on the soil surface in terms of quantity and species composition, as well as variety of forms.

Among the paremiological units collected in this section, those formed on the basis of the names of insects were identified, and in the course of research they were divided into groups of 18 species, as well as the linguistic bases that served as the origin of each of them. paremes were determined by sequence.

In particular, in Uzbek folk proverbs the lexeme flea is used for the following reasons:

1. In appearance: "small".

Burga tutmoqqa ham barmoq ho'llamoq kerak.

This proverb says that even to achieve a small result, you must always try and work.

2. By way of life: "resident".

Bit - g'amdan, Burga - namdan, Pashsha - dimdan, Kana - go'ngdan.

3. According to biological characteristics: "blood-sucking."

Burgaga achchiq qilib, Koʻrpaga oʻt qoʻyma. Burgani deb poʻstinni olovga tashlama.

A feature of the flea insect is blood sucking, as a result of which a person feels uncomfortable and ill. In order to get rid of this situation, it is recommended not to give up blankets or furs, but to eliminate this situation by fighting the pest itself.

4. By movement: "fast moving", "crawling".

Burga qochar oyoqqa, Bit qoladi tayoqqa. Burga ketdi sayoqqa, Bit qoldi tayoqqa. Burga sakraydi, bit yoʻrgʻalaydi. It achchigʻini turnadan olar, Bit achchigʻini burgadan.

Observations show that most of the Uzbek folk proverbs formed on the basis of the lexeme flea were created on the basis of its movement.

Standards for the design of linguistic corpora

Until now, we have talked about the properties of the case in abstraction from specific technological solutions that make it possible to implement them in practice. Such solutions can be different, and the more markings the case contains, the more application programs are developed for its operation, the more diverse and difficult to compatible the technical solutions can become. The incompatibility of standards used by corpus creators in different countries and research centers threatens the possibility of widespread data exchange, unification and mutual enrichment of corpora, which is so important for corpus linguistics.

British National Corpus (BNC)

One of the most famous and popular corpora of the English language (but far from the only one) is the British National Corpus (BNC). This corpus was created through the joint efforts of several British universities and publishing houses, as well as the British Library, between 1991 and 1994. The corpus includes written and spoken texts in British English from the late 20th century, belonging to a wide variety of genres and functional styles. The corpus is fragmentary: texts of more than 45,000 words are presented in excerpts (which makes it possible to avoid the influence of the individual style of a particular author on the overall results).

The total volume of the corpus is slightly more than 100,000,000 word usages. BNC texts are marked up in the SGML standard in accordance with TEI recommendations.

The BNC corpus is equipped with morphological markings: each word form is characterized by its belonging to the part of speech, the category within the part of speech and the form of inflection. This marking was carried out automatically, which led to errors in 1.7% of cases, and 4.7% of word forms could not be unambiguously interpreted and received a "double morphological code". A fragment of the corpus, constituting 2% of its total volume, was selected for more detailed ("manual") morphosyntactic marking.

Operation of the housing is carried out using a number of specially created SGML processing programs. Limited access to corps resources is available free of charge via the Internet <http://www.natcorp.ox.ac.uk/>, however, in order to take advantage of all its capabilities, you must purchase a CD-ROM or register for a fee for on-line access.

BNC data is widely used in the compilation of dictionaries, grammars and textbooks of the English language, in linguistic research, in work on artificial intelligence, as well as in the practice of teaching English.

FRANTEXT

One of the historically first and largest electronic collections of texts today is the French Frantext database. Strictly speaking, this is not a corpus, but the system of its operation allows the researcher to form his own "working corpus" taking into account a number of parameters (author, date, genre, size, etc.).

As already noted, work on creating the database began in 1957 as part of the preparation of the 16-volume "Thesaurus of the French Language," but over time, replenishment and development of means for operating the corpus became an independent task. Large financial resources were invested in the creation of Frantext: an entire laboratory of the French National Center for Scientific Research (CNRS), consisting of 30 to 50 people, worked on it for almost half a century. Currently, Frantext contains 3,737 texts from the 16th – 20th centuries. (about 210,000,000 uses of words) and continues to be constantly updated. The bulk (about 80%) consists of literary texts, but it also includes scientific and technical works. A little more than half of the texts in the database (1940 texts, 127,000,000 word usages) are provided with morphosyntactic markings [5].

External access to Frantext has been open since 1992 for corporate users (libraries, universities, etc.) and is paid. Free access is provided to the bibliographic database and to the electronic version of the Thesaurus of the French Language (TLFI).

In recent years, work has been carried out to deepen the "historical perspective" of Frantext: databases of texts from the Old French (IX – XIII centuries) and Middle French (XIV – XV centuries) periods have been added to it, and anyone can use these databases for free.

To some extent, the advantages of Frantext - its colossal size and long history of formation - are at the same time the source of its problems. Developed in the 60s - 70s. formats and operating systems are currently very outdated and do not meet the capabilities of modern technology and the needs of researchers. Modernization of Frantext - in particular, its translation into the XML standard and markup in accordance with the TEI recommendations - is a complex task, and at present it is difficult to say when it will be solved.

In recent years, especially in student coursework and dissertations, examples often appear, the source of which is defined as the "Internet". Such a practice is unacceptable in scientific research, since the World Wide Web itself can be considered as a corpus of texts to an even lesser extent than virtual libraries. Without a clear indication of the source of the example and a definition of its functional, stylistic and genre affiliation (despite the fact that, as far as we know, a "genre classification" of Internet texts has not yet been developed), it is impossible to assess the linguistic status of the fact illustrated by the example. At the same time, the Internet certainly represents a new and extremely interesting environment for the existence of language with its unique genres ("chat rooms", "forums", electronic correspondence, entries in guest books, etc.), which deserves the closest attention of linguists.

4. Conclusion

The creation and development of a wide variety of text corpora in different languages – both "large" and endangered – can rightfully be considered one of the priority tasks of linguistics. These corpora will provide future generations of researchers with a reliable and easily accessible source of data on

the functioning of the language in a wide variety of areas and on the culture of the people speaking this language. When creating text corpora, one should be guided by international standards and recommendations designed to ensure the safety and accessibility of data regardless of changes in technology and software. Uzbek folk proverbs, formed on the basis of the lexical-semantic group "insect," occupy a significant place in the expression of positive, negative and neutral meanings. This is explained, firstly, by the fact that insects have more harmful aspects than beneficial ones, and secondly, in relation to those that have beneficial properties (bees, silkworms, leeches), there is a relatively large number of "pests" among them (lice, can be explained by the abundance of fleas, butterflies, moths, ticks, flies, scorpions, mosquitoes).

The results of the study of Uzbek folk parems, formed on the basis of the lexical-semantic group "insect", show that the insects used in them do not represent only the same seme. Depending on the speech situation and the actual possibility, their meaning may vary.

References:

1. British National Corps: <http://www.natcorp.ox.ac.uk/> FRANTEXT Corps: <http://www.atilf.fr>
2. Brief history of SGML development:
3. <http://www.sgmlsource.com/history/sgmlhist.htm>
4. Machine Fund of the Russian Language: <http://www.irlras-cfrl.rema.ru/> Russian National Corpus: <http://www.ruscorpora.ru>
5. World Wide Web Consortium (XML specification): <http://www.w3c.org> Text Encoding Initiative (TEI): <http://www.tei-c.org>
6. Corpus Encoding Standard (CES): <http://www.cs.vassar.edu/CES
7. Toirova G., Astanova G., Rahimova N. Artistic Expressions of a Situational Pragmatic System. // International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-3, September 2019. –P.4591-4593
8. Toirova G.,Yuldasheva M., Elibaeva l. Importance of Interface in Creating Corpus. // International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-2S10, September 2019. –P.352-355.
9. Toirova G, Abdurahmonova N., Ismoilov A., Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. 2022 7th International Conference on Computer Science and Engineering (UBMK) Sep. 14 - 16, 2022, Diyarbakir /Turkey pp. 73–75.
10. Зимовец Н.В. К вопросу о происхождении английских пословиц и поговорок / Актуальные вопросы переводоведения и практики перевода. – Россия. Г. Белгород. 2013. – С. 112–118.
11. Сирот И.М. Русские пословицы библейского происхождения.–Брюссель:ЖизньсБогом,1985.– 128с философ.(PhD)по филол.наук.–Ашхабад, 2018.–24с.
12. Basharan S. HadislerinTurk Atasozlerine Tesiri.–Turkiye, 2017.–22 b.
13. Джураева Б.М. Лингвистические основы и прагматические особенности формирования узбекских народных пословиц: Дисс. д.ф.н. – Самарканд, 2019. – 237 с..
14. Тошева Д.А. Лингвокультурологическая характеристика пословиц с зоонимическим компонентом: Филол.фан. Доктор философских наук (PhD) дисс... - Т., 2017. - 134 с.