Synergy: Cross-Disciplinary
Journal of Digital Investigation
(ISSN 2995-4827) VOLUME 02 ISSUE 8, 2024

THE ISSUE OF CORPUS DESIGN IN THE CREATION OF ELECTRONIC DICTIONARIES

Latipova Gulasal Bahrom qizi
Doctoral student of Alisher Navoi Tashkent State University of Uzbek Language and Literature

Abstract:

This article reflects the history of lexicology, lexicological scientists and their scientific researches, differences and similarities between modern and ancient lexicology, periodization of lexicology. In particular, the causes and factors of the emergence of lexicology are listed.

Keywords: *lexicography, principle, chronology, classification, acceleration, social task, phraseological unit.*

- If the initial lexicographic studies are analyzed, it is appropriate to note that samples from different sources obtained for one or another word were classified in the form of a card file, and their classification at a certain level was used as a lexicographic principle. Now, this arduous and time-consuming process has been further improved by the scientific advances of corpus linguistics.
- In world linguistics, the possibilities of the corpus for lexicographical processes are recognized by scientists.
- About the four main types of lexicographic corpus, E. Walter states the following:
- In the 3rd edition of the Longman Dictionary of Modern English, information based on the transcription of spoken texts from English-language television channels was used to create a dictionary. In this dictionary, words and phrases taken from oral and written texts are analyzed graphically in a comparative aspect;
- The use of the corpus is also of practical importance to describe the diversity of the regional language. It is used in the analysis of commonalities and differences in the specific aspects of language and speech phenomena of English speakers in Britain, Great Britain and Australia;
- It should be taken into account that the vocabulary created on the basis of corpus analysis is in synchronic or diachronic aspect. Synchronous information is created through recently created sources in the creation of modern dictionaries. The diachronic corpus was created on the basis of diachronic sources and serves to identify lexical units that have entered the archaic layer today. Such corpora are "monitor corpora", that is, corpora that are equally distributed in all

respects, and differ in that this type of information is constantly updated if the main corpora is stable.

- Corpus types differ to a certain extent when creating a dictionary, depending on whether the speakers have a local language or a non-local language. Such corpora are useful in creating educational dictionaries, and a corpus is created based on various written works, that is, essays, recorded during the educational process of representatives of other nationalities.

For lexicographers, the types of genres in a corpus's content are highly relevant. Whether the vocabulary is general or specialized depends on the content taken for the sample. Therefore, in lexicographic research, preliminary processes are carried out on the basis of these goals in order to create a corpus. For example, if a lexicographer wants to create a sectoral educational dictionary related to the field of tourism, attention is paid to the complete formation of the corpus of texts related to this field. If translation is dealing with dictionaries, it is certainly appropriate to use a parallel corpus with segmented translation alternatives. It is necessary for the translation dictionaries to include all the possibilities of the language, especially in revealing the polysemantic features of each lexeme, with a good understanding to which layer this or that word belongs, to give its explanations and examples, and to include the texts for the appropriate branched corpus in a large amount of resources.

Together with Adam Kilgarriff and his Lexical computing Ltd (Lexical computing Ltd), Sketch Engine technology was created. This multilingual platform is the most important linguistic resource for lexicographers to create a dictionary based on data from web corpora through concordances.

A unique aspect of corpus-based computer lexicography is that it has enabled the creation of a linguistic model to digitize the linguistic resource in the field of computational linguistics and NLP, thereby modeling natural language capabilities. For example, in artificial intelligence-based translation technologies such as machine translation, machine learning has been implemented by using large corpora to match the linguistic rules of languages and their semantics on a multilingual platform.

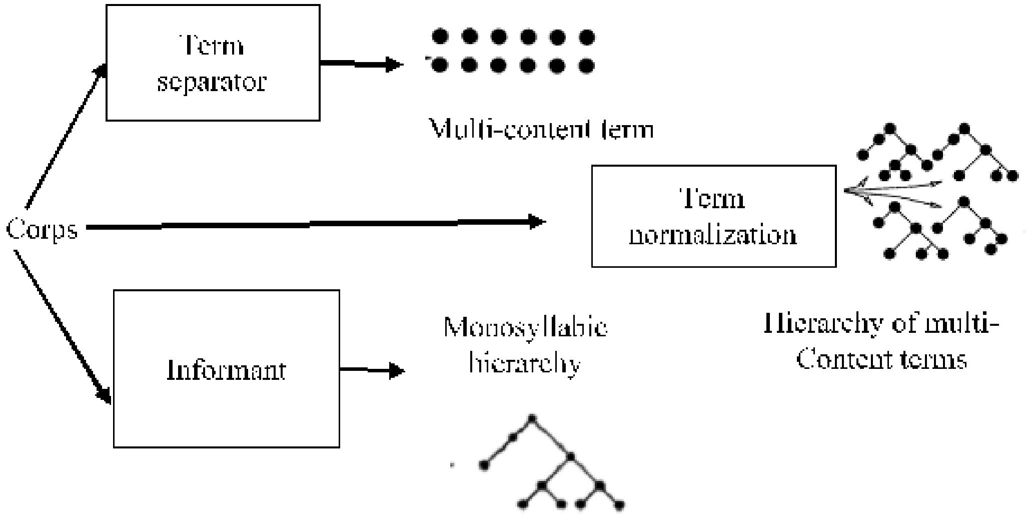
Linguee was founded in 2007, and based on bitext, i.e., translated parallel texts on sites, the dictionary interface extracts translation units in the context of the CRAWLER technology. This translation dictionary platform is integrated with DeepL translation technology. This means that using the vast possibilities of the corpus, it is possible to solve a number of language-related problems not only in lexicography, but also in the field of computer technology and NLP.

One of the electronic dictionaries formed on the basis of another parallel corpus is Tatoeba. Notably, this platform includes a parallel corpus of all world languages. The Uzbek language is also included in this platform, so the base is not that large.

In computer lexicography, the phenomenon of homonymy is of particular importance. In particular, F. Atsushi is known in science as a scientist who has conducted many years of scientific observations on the identification of polysemous or homonymous words based on the corpus. According to him, it is possible to solve homonymous or polysemous words using a similarity based model. According to F. Atsushi, this system uses a manually annotated database for each sentence to determine the meaning of words. Based on this, the scientist proposes three main methods for solving verbal polysemy: 1- determining the weight value by calculating the similarities of each word; 2- manually check the samples selected for the system in small quantities; 3- effective use of previously calculated similarities as a base in subsequent search stages.

In addition, other foreign scientists have also carried out research on linking the semantic relations of words in computer lexicography based on the corpus. Relatively early scientific observations about the automatic creation of hyporonymic links by parsing texts (syntactic analysis) can be found in the works of M. Hearst. In his works, the problem of automatic separation of hyponymic lexical relations from unconstrained text was solved. It is possible to form a thesaurus structure created by a large amount of manual labor by creating an opportunity for the researcher to determine the relationship of the words found in the texts through lexical-syntactic frames. At the same time, the scientist emphasizes that the machine readable dictionary is important, but it is appropriate to refer to textual resources when giving their meaning. In our opinion, this approach was able to demonstrate the need for corpus-based linguistic analysis in time.

In works carried out by scientists such as Bourigault (1993), Justeson and Katz (1995), Daille (1996), Morin (1999). point of view prevails. According to it, classes are created by clustering words in similar contexts. Connections between words are distinguished according to predicative signs built on cause-and-effect relationships. M. Emmanuel shows that there are two main directions of corpus-based terminology in this regard: 1) definition of a term through alternative terms; 2) automatic creation of a thesaurus through a semantic relation connected to the terminological database. The scientist presents the following system for hierarchical design of words:



In his view, special conceptual link can be found corpus-based lexical-syntactic frames through hyperonyms, and conceptually linked term pairs can be identified through the lexical-syntactic frame reference base.

It is worth noting that the Uzbek language has not yet compiled an electronic linguistic dictionary, which is classified according to semantic categories. However, a huge number of types of linguistic dictionaries are created in the form of books, concentrating them on one base precisely on a thesaurus, presenting the SEMAS of various meanings of one word or another according to their semantic content is considered one of the pressing issues facing computer lexicography.

REFERENCES:

1. Hasanov B. “Javohir xazinalari. Qo‘lyozma lug‘atlar”, T.: G‘afur G‘ulom n. 1989-y.
2. “O‘zbek tili leksikologiyasi”. T.: Fan n. 1981-y.
3. Hamidov Z. “Yozuv va majoziy ma’nolarni sharhlovchi lug‘at”. O‘TA jurnali. 1999-yil, 1-son.
4. Munday J. Corpus-Based Translation Studies Research and Applications (Edited by Alet Kruger, Kim Wallmach and Jeremy Continuum Advances in Translation Studies–P.19
5. Jeremy Continuum Advances in Translation Studies - P. 16.
6. Blum-Kulka, Shoshana and Eddie A. Levenston (1983) ‘Universals of Lexical Simplifi cation’, in Claus Faerch and Gabriele Kasper (eds) Strategies in Interlanguage Communication, London:

Longman, 119–39.; Shlesinger, Miriam (1989) ‘Simultaneous Interpretation as a Factor in Affecting Shifts in the Position of Texts in the Oral-Literate Continuum’. Unpublished MA Dissertation, Tel Aviv, Tel Aviv University.;Toury, Gideon (1985) ‘A Rationale for Descriptive